DWD Normalization of Micro-Array Batch and Cross-Platform Effects

J. S. Marron

Dept. of Statistics and Operations Research

June 28, 2004

Main Lessons

- Normalize data to “make comparable”
  - For source and batch effects
  - Across platforms
  - Based on “DWD” (Distance Weighted Discrimination)
- Allows combining data sets
  - Bigger data sets → More statistical power
  - Set your data among larger caBIG data base
- Visualization is crucial
  - To see “why it works”
  - As diagnostics to understand and handle failures

Website: Details & More Views

DWD caBIG Web Page:

http://genome.med.unc.edu:8080/caBIG/DWDindex.htm

- Many more “steps”
- Also Clustered Tree View Heat Map Views

Key Philosophical Point

- Competing Paradigms:
  - Visually: what do we look at?
  - Conceptually: how do we think?

Gene by Gene

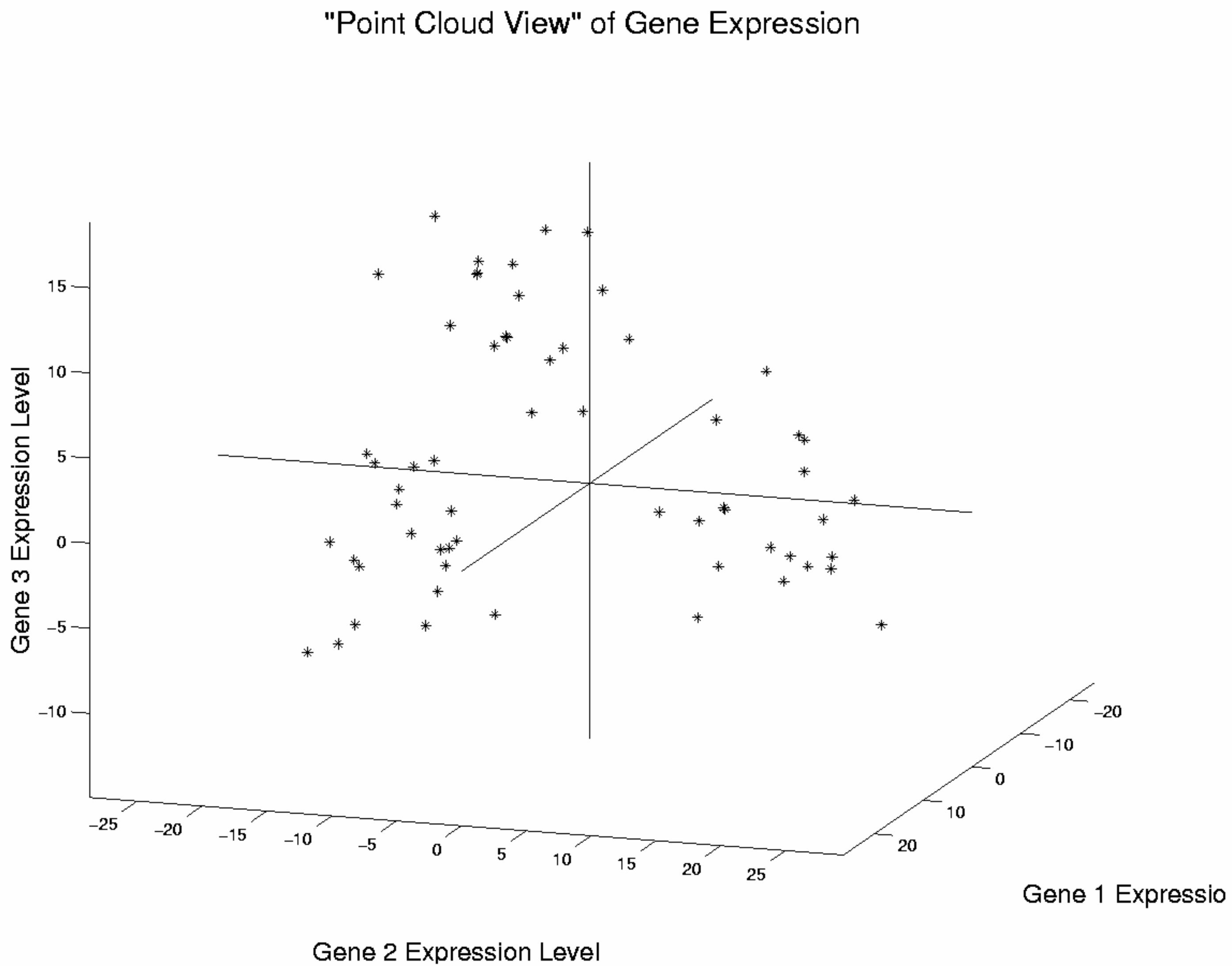
vs.

Multivariate “point cloud”

- Will illustrate power of multivariate concept
  - While showing to combine data across platforms

UNC Lineberger

Illustration of Multivariate View: Raw Data



UNC Lineberger

Illustration of Multivariate View: Highlight One

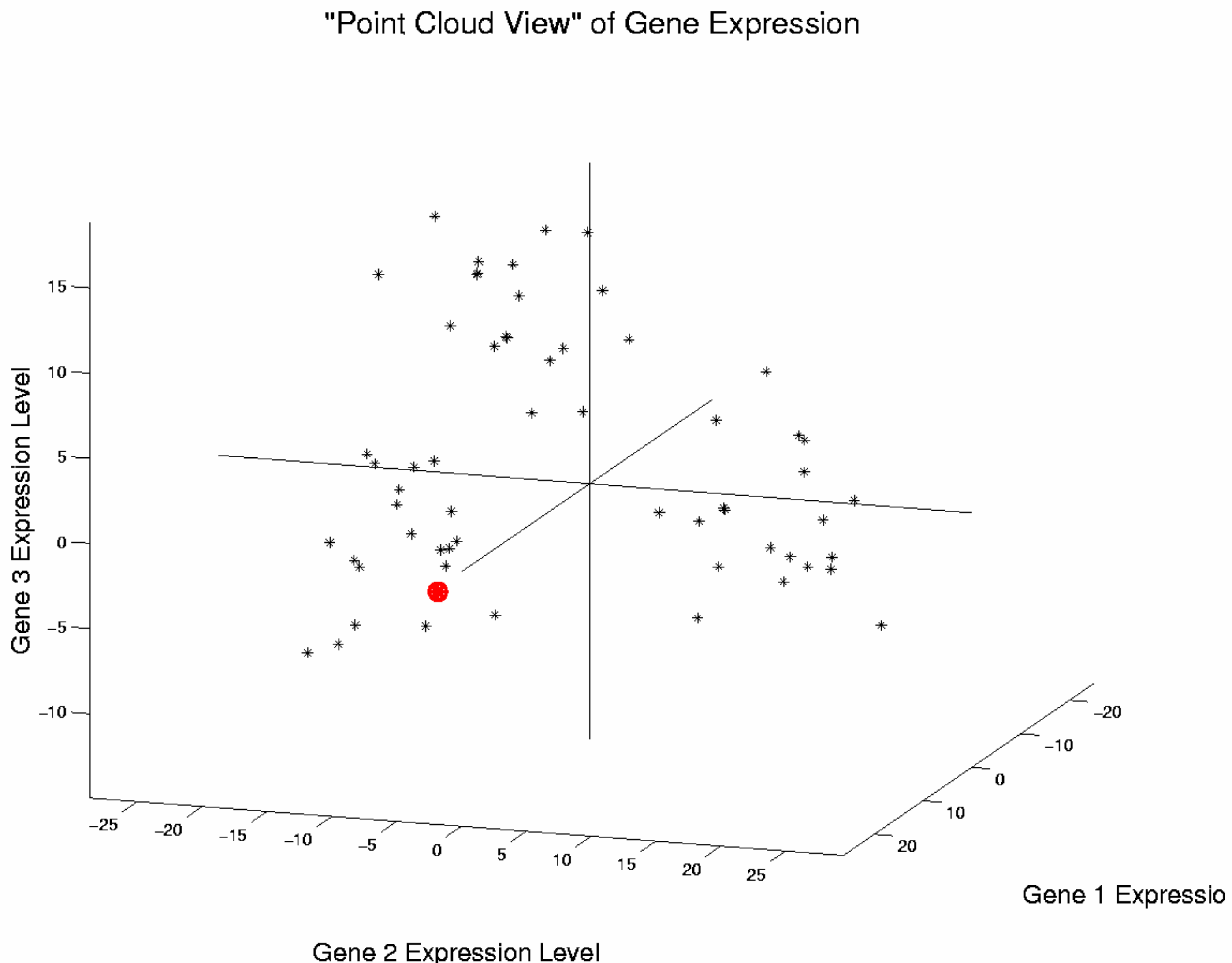
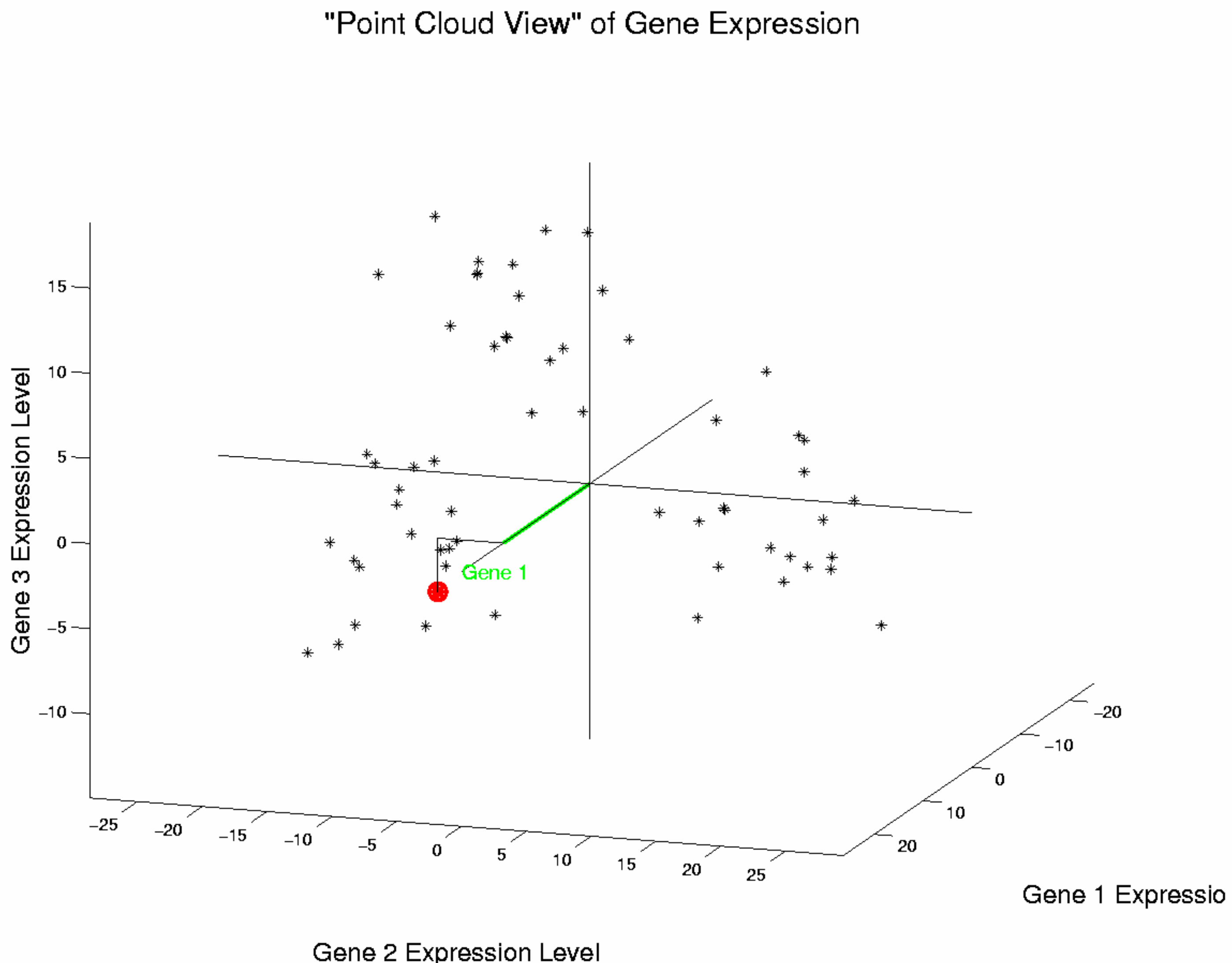


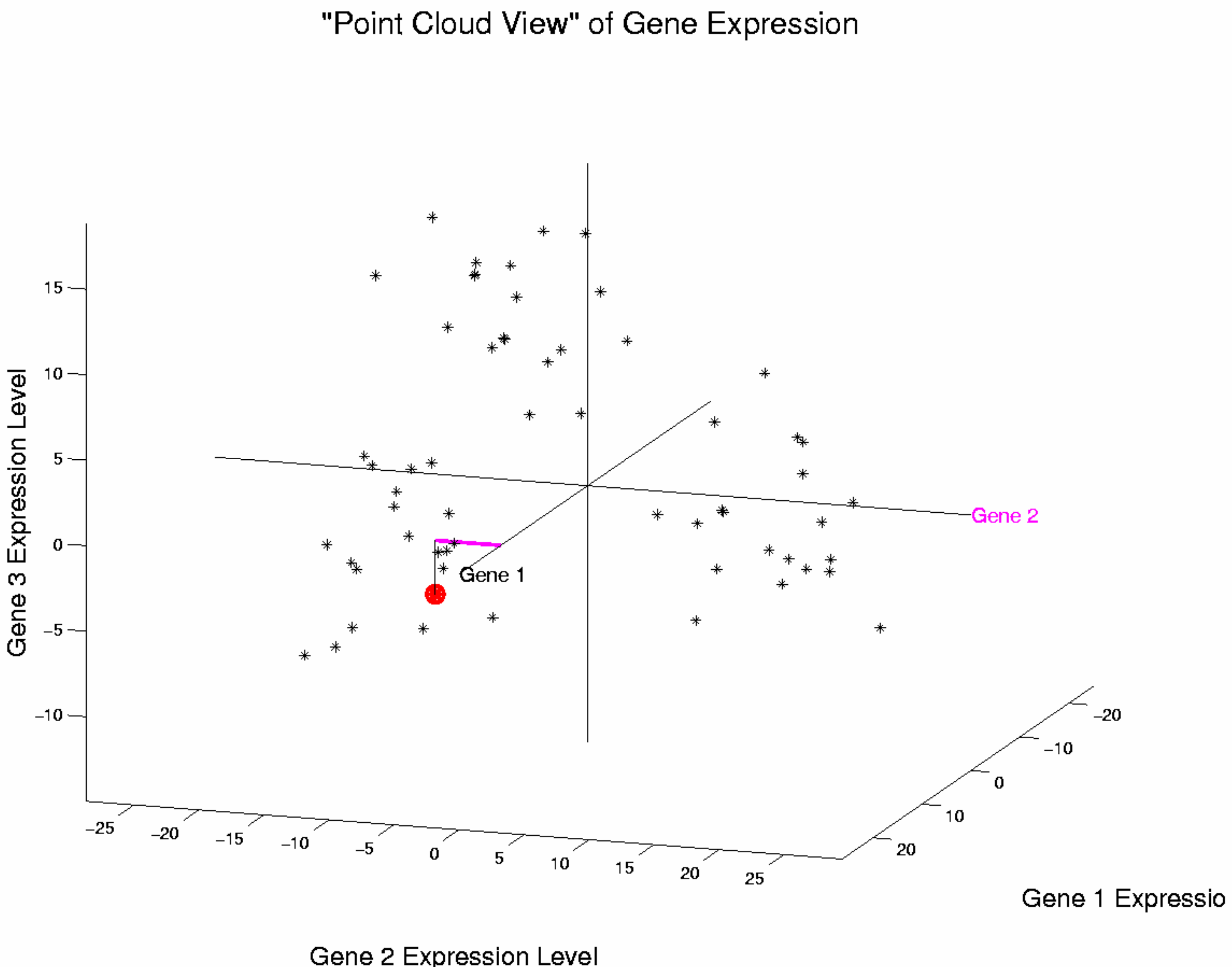
Illustration of Multivariate View: Gene 1 Express'n

UNC Lineberger



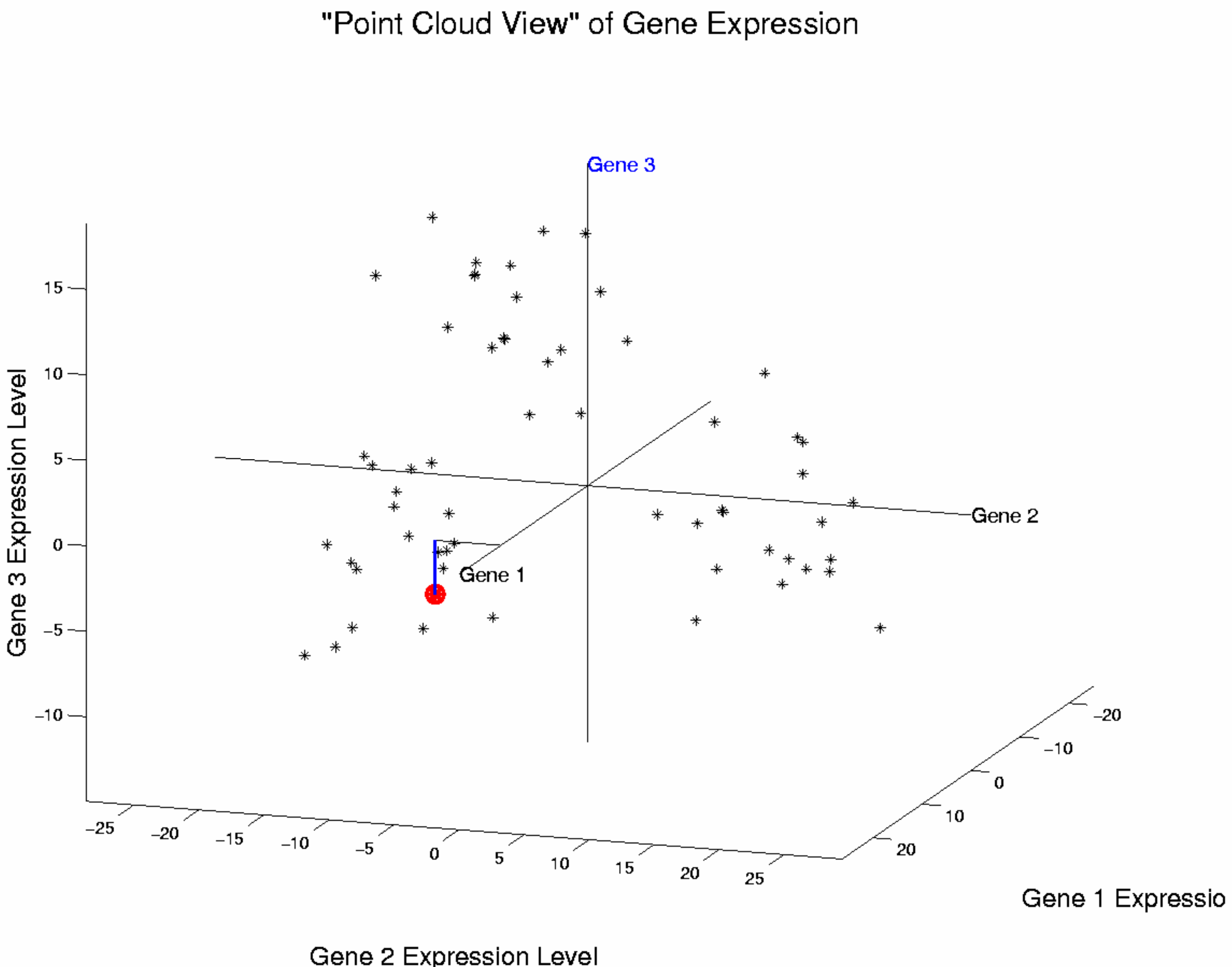
UNC Lineberger

Illustration of Multivariate View: Gene 2 Express’n



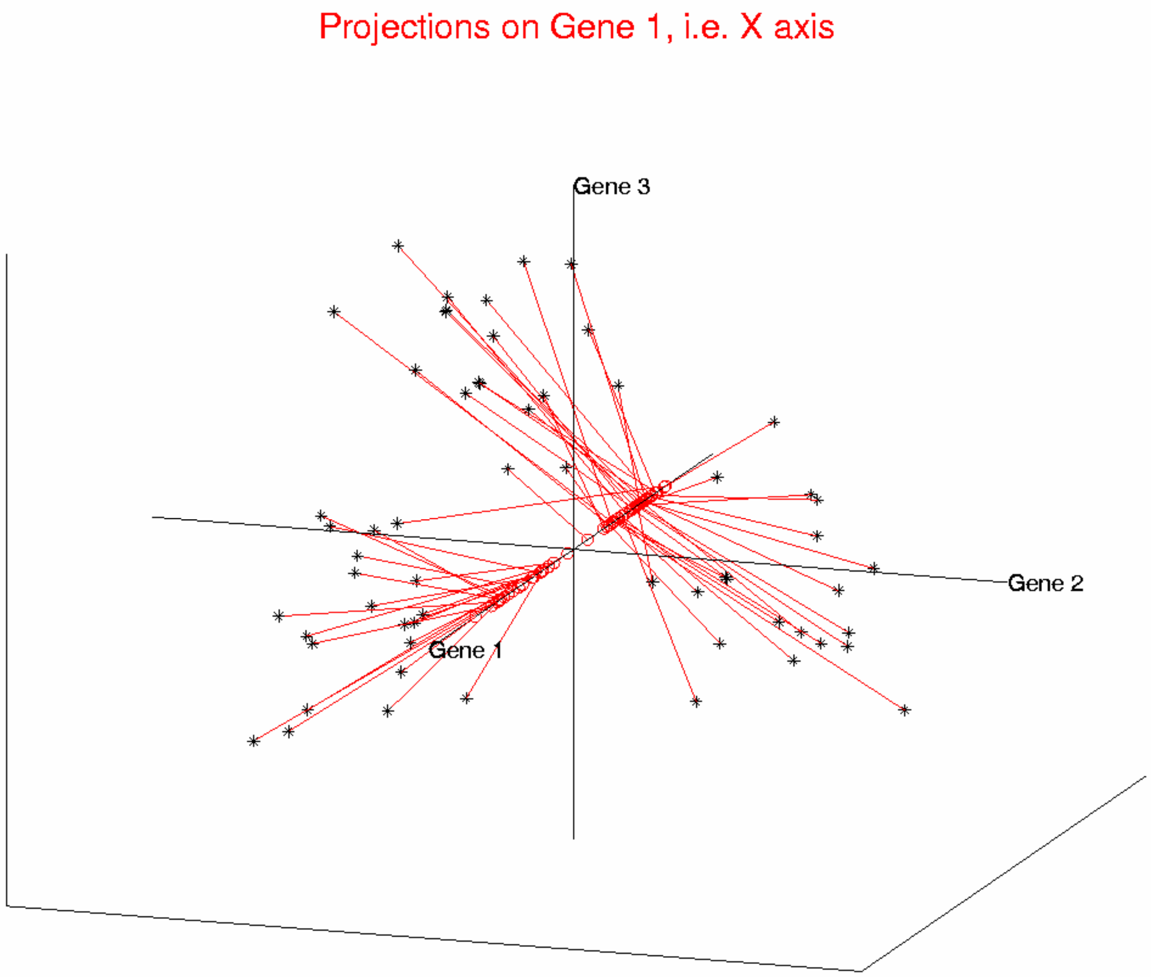
UNC Lineberger

Illustration of Multivariate View: Gene 3 Express’n

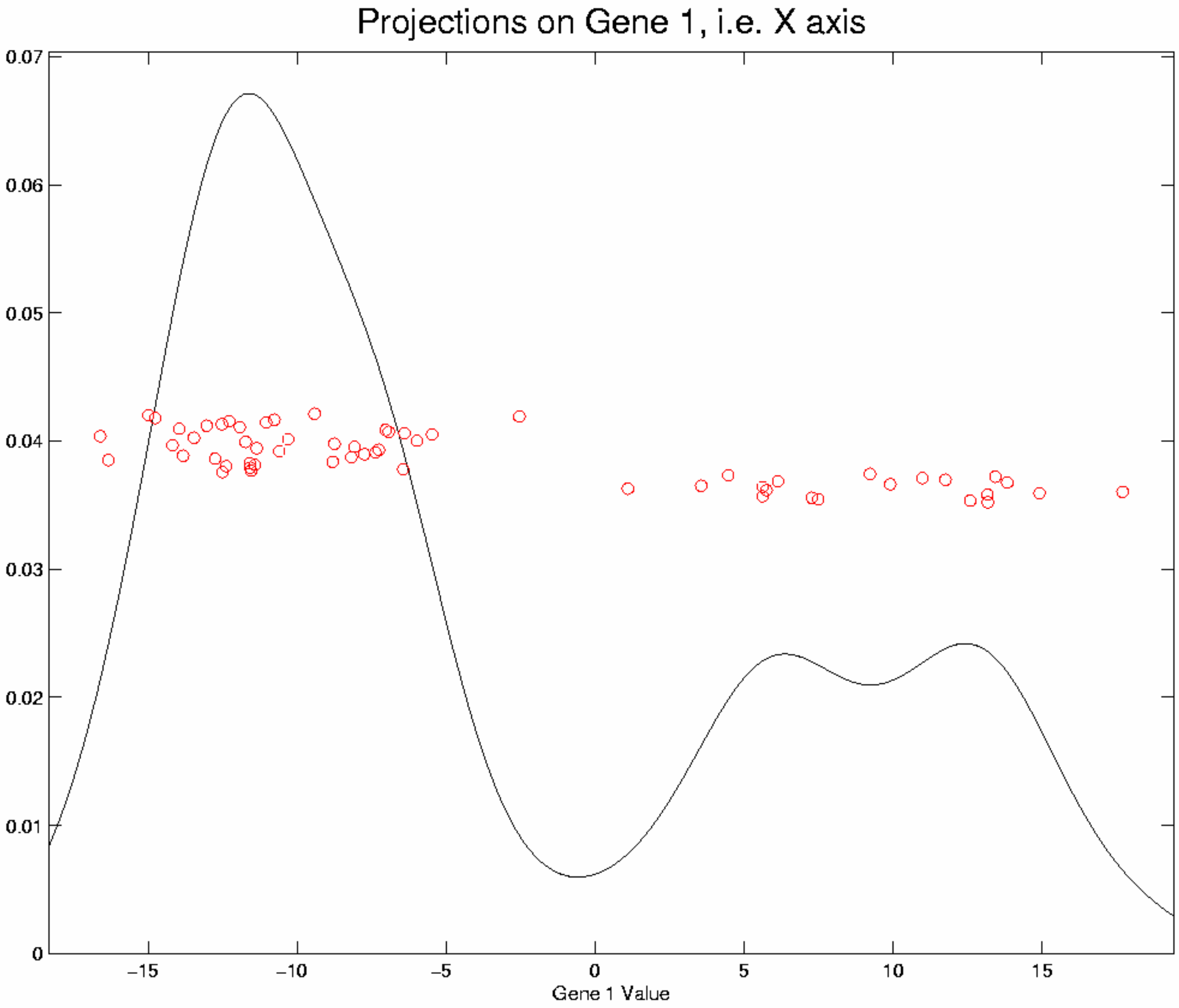


UNC Lineberger

Illust’n of Multivar. View: 1-d Projection, X-axis

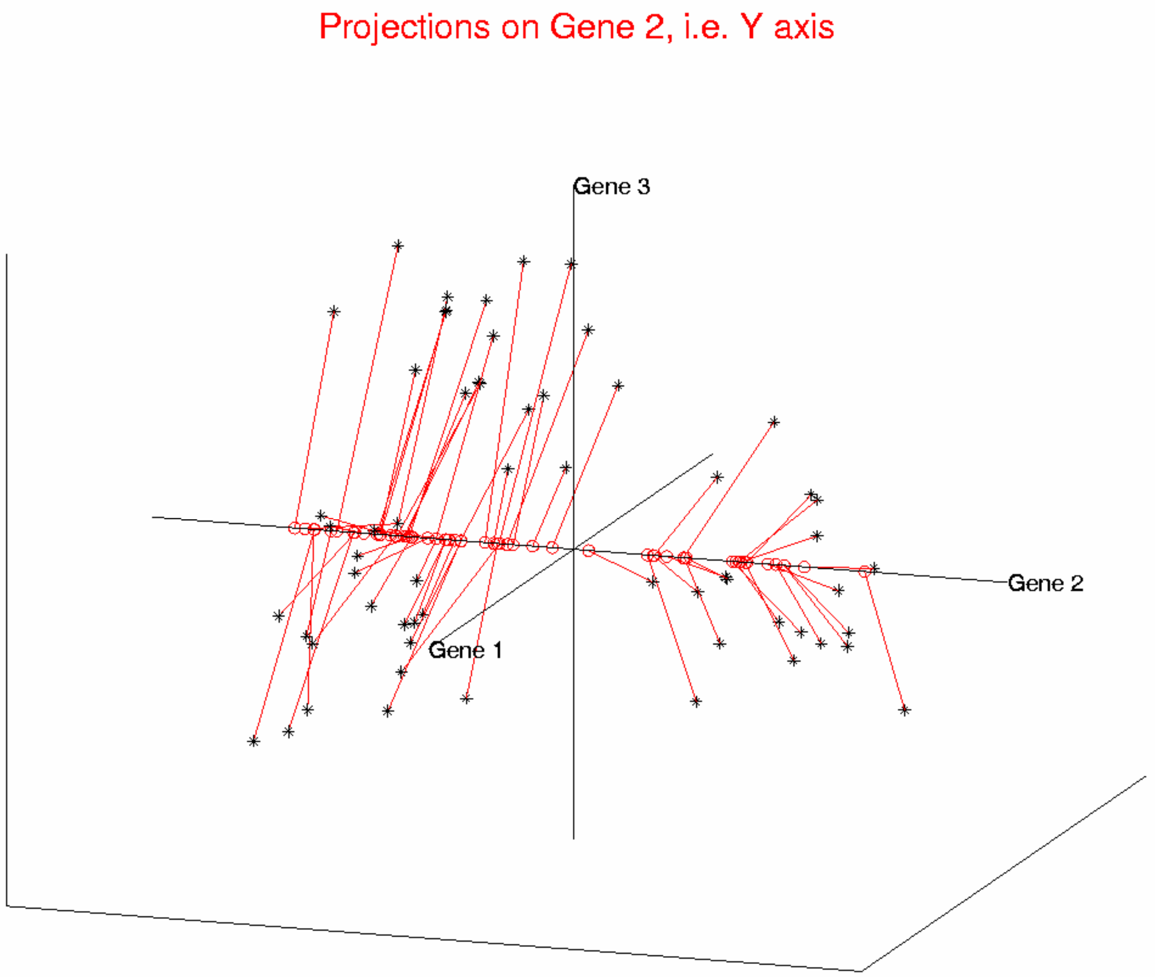


Illust'n of Multivar. View: X-Projection, 1-d view

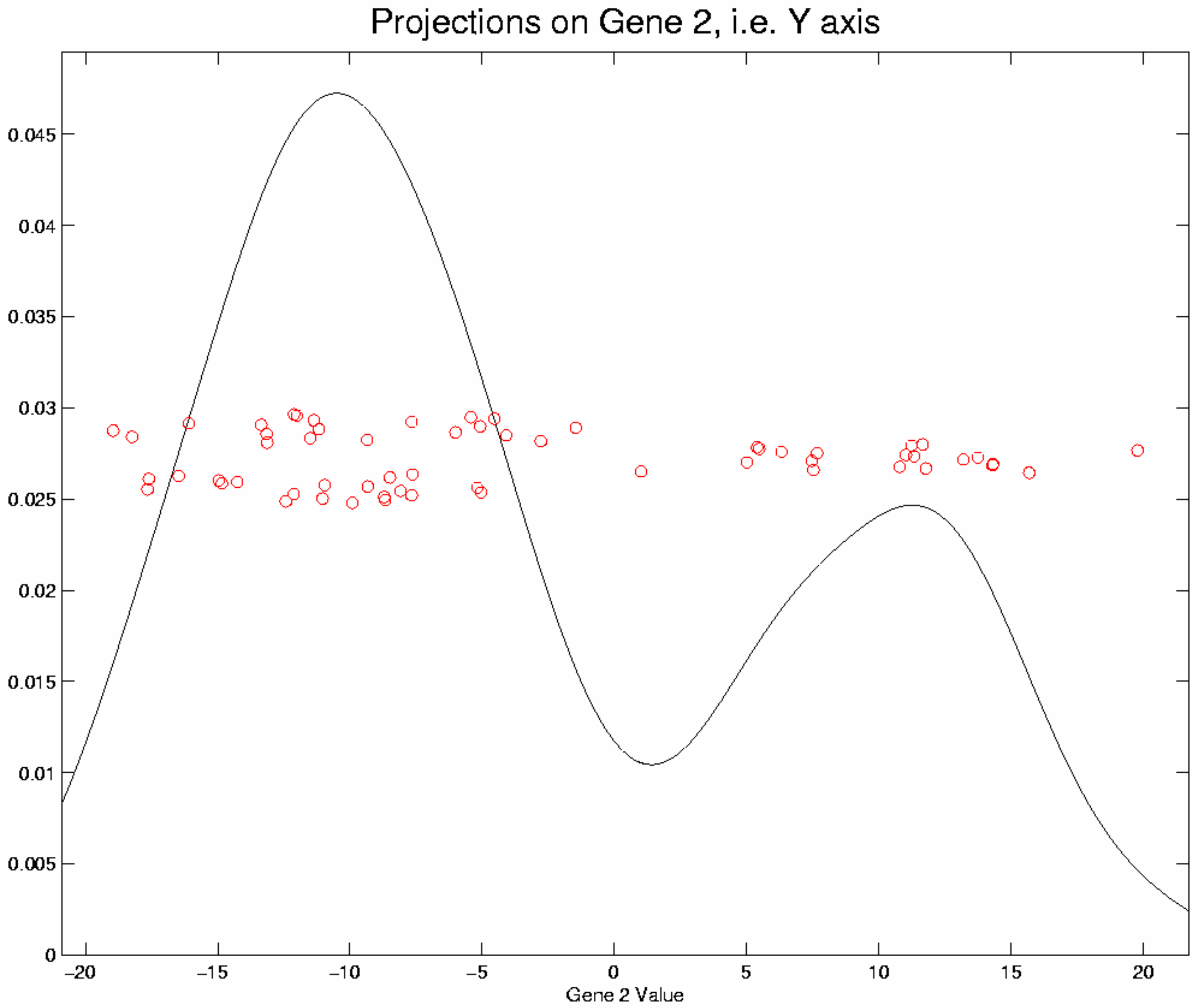


UNC Lineberger

Illust’n of Multivar. View: 1-d Projection, Y-axis

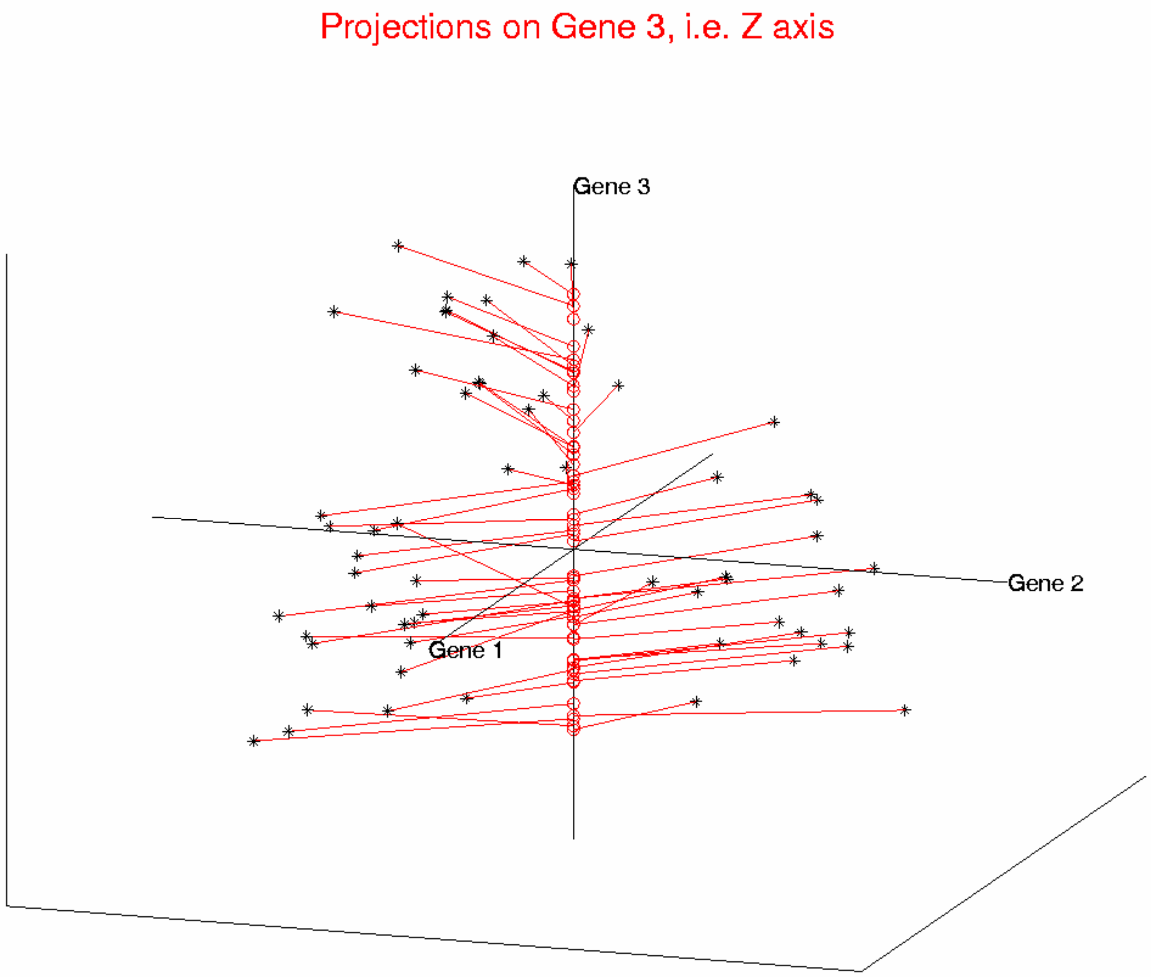


Illust'n of Multivar. View: Y-Projection, 1-d view

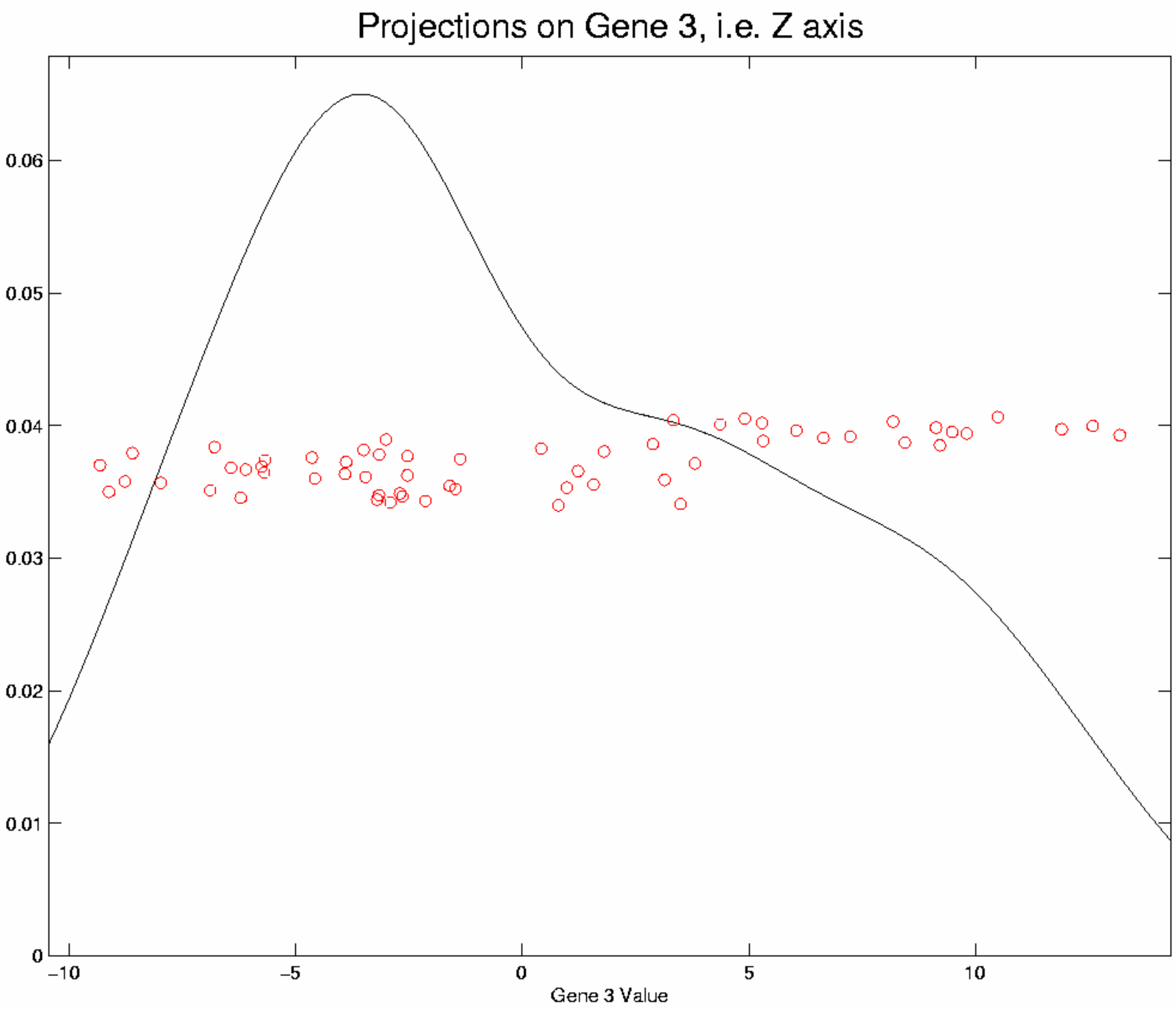


UNC Lineberger

Illust’n of Multivar. View: 1-d Projection, Z-axis

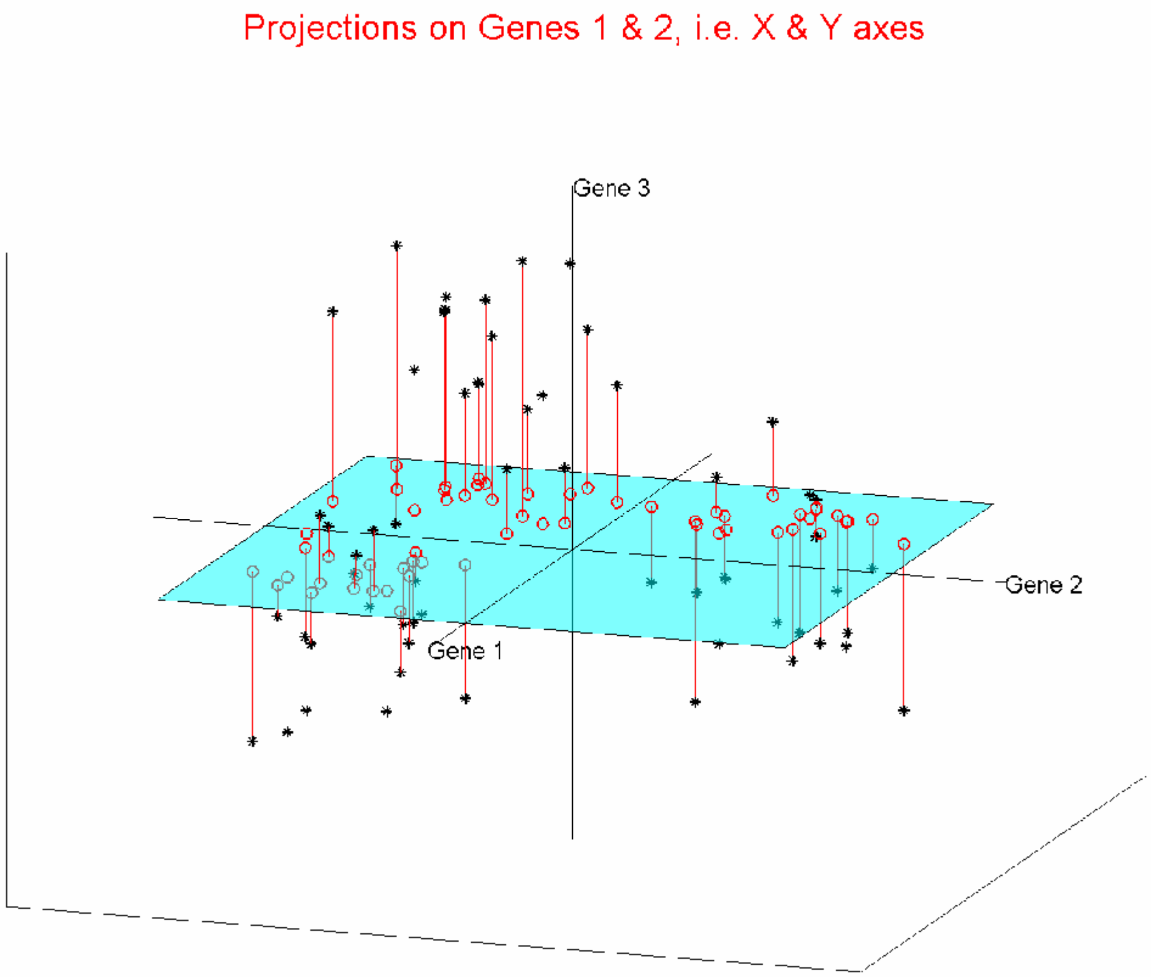


Illust’n of Multivar. View: Z-Projection, 1-d view



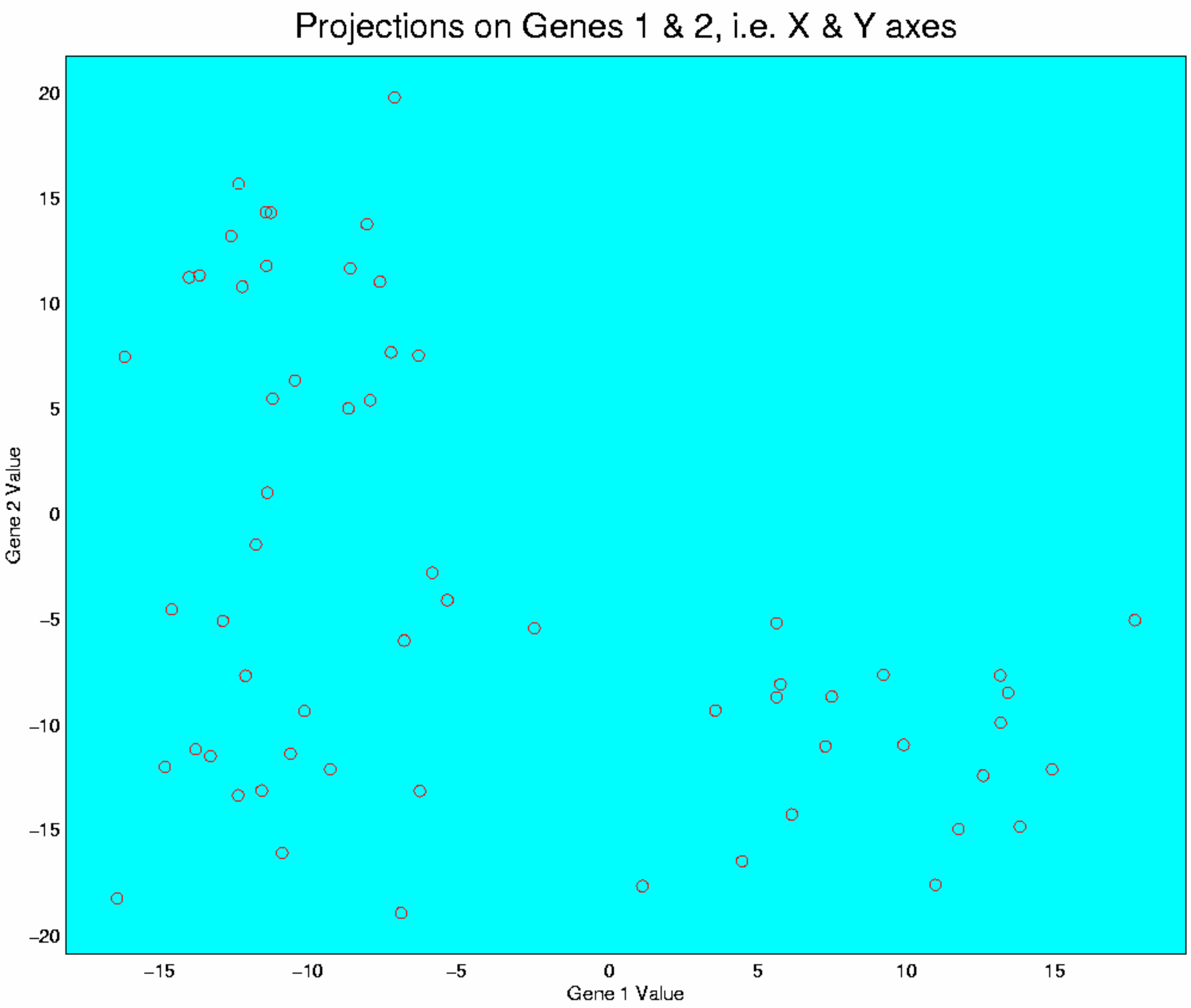
UNC Lineberger

Illust’n of Multivar. View: 2-d Proj’n, XY-plane



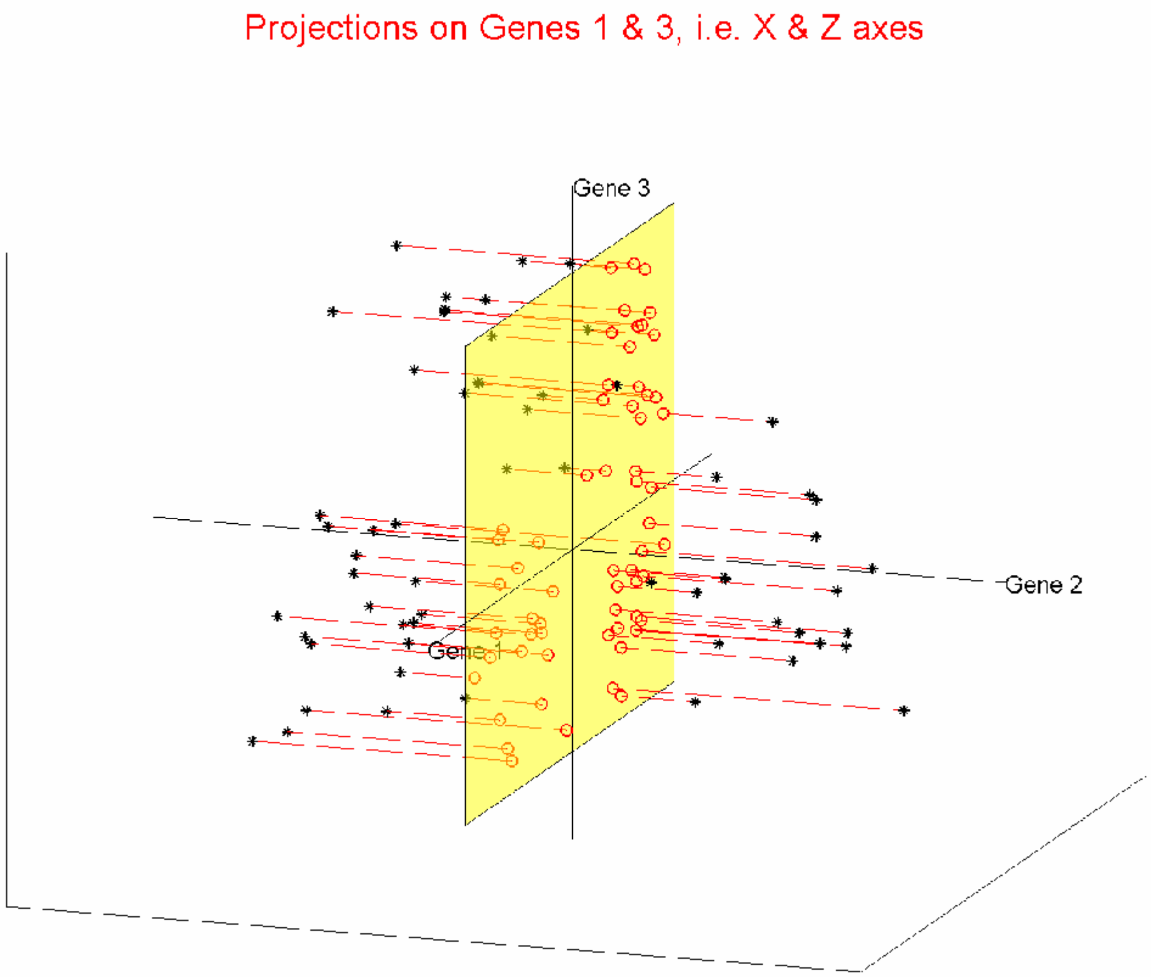
UNC Lineberger

Illust’n of Multivar. View: XY-Proj’n, 2-d view

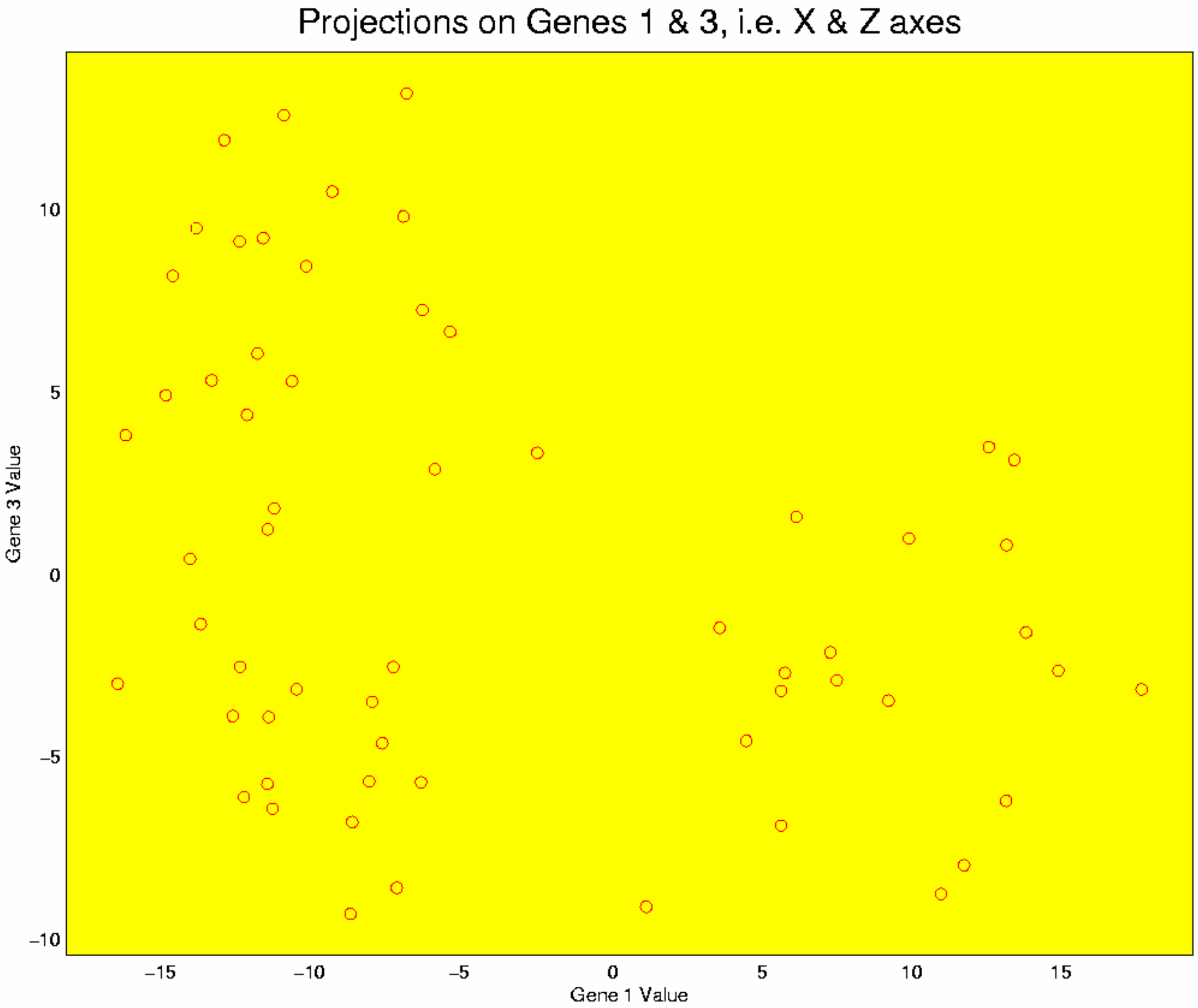


UNC Lineberger

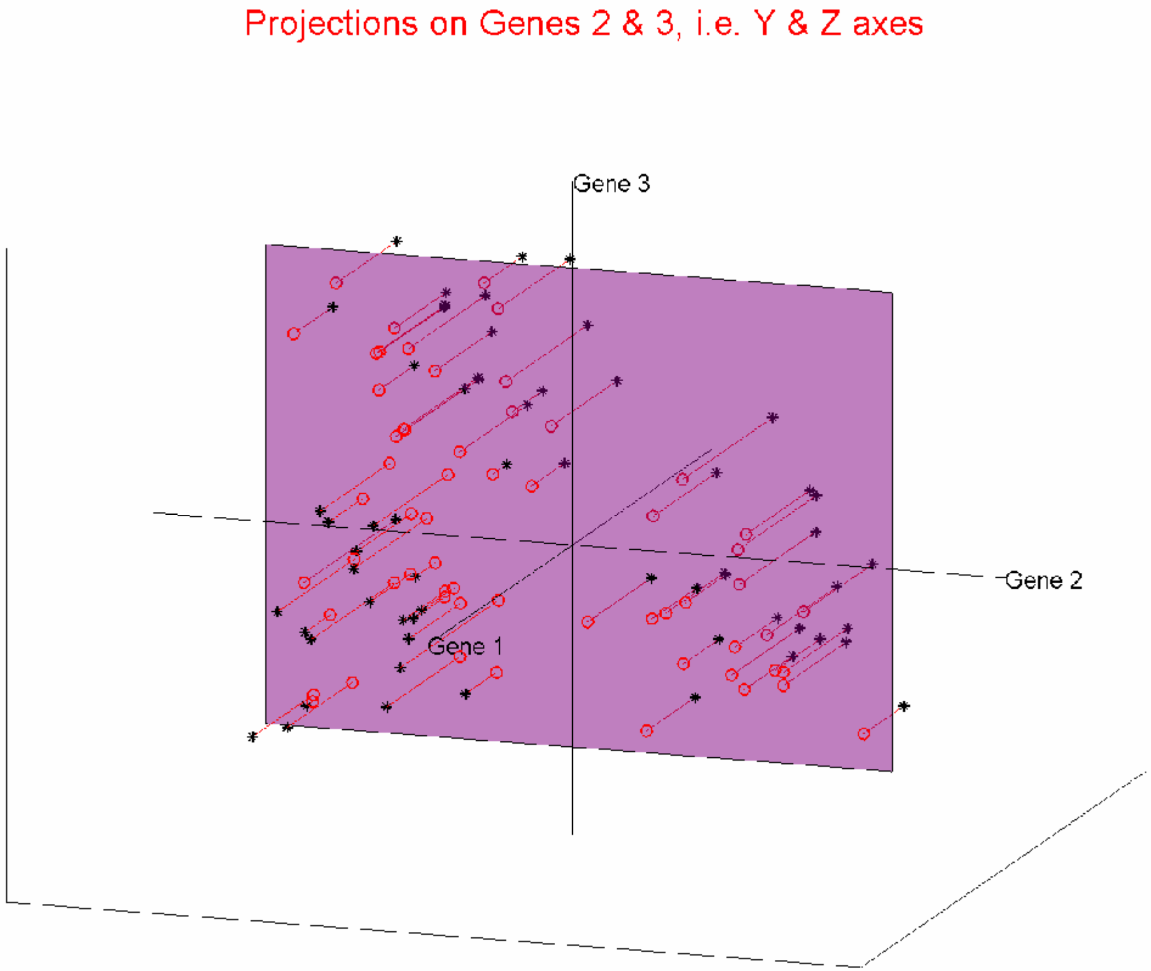
Illust'n of Multivar. View: 2-d Proj'n, XZ-plane



Illust’n of Multivar. View: XZ-Proj’n, 2-d view

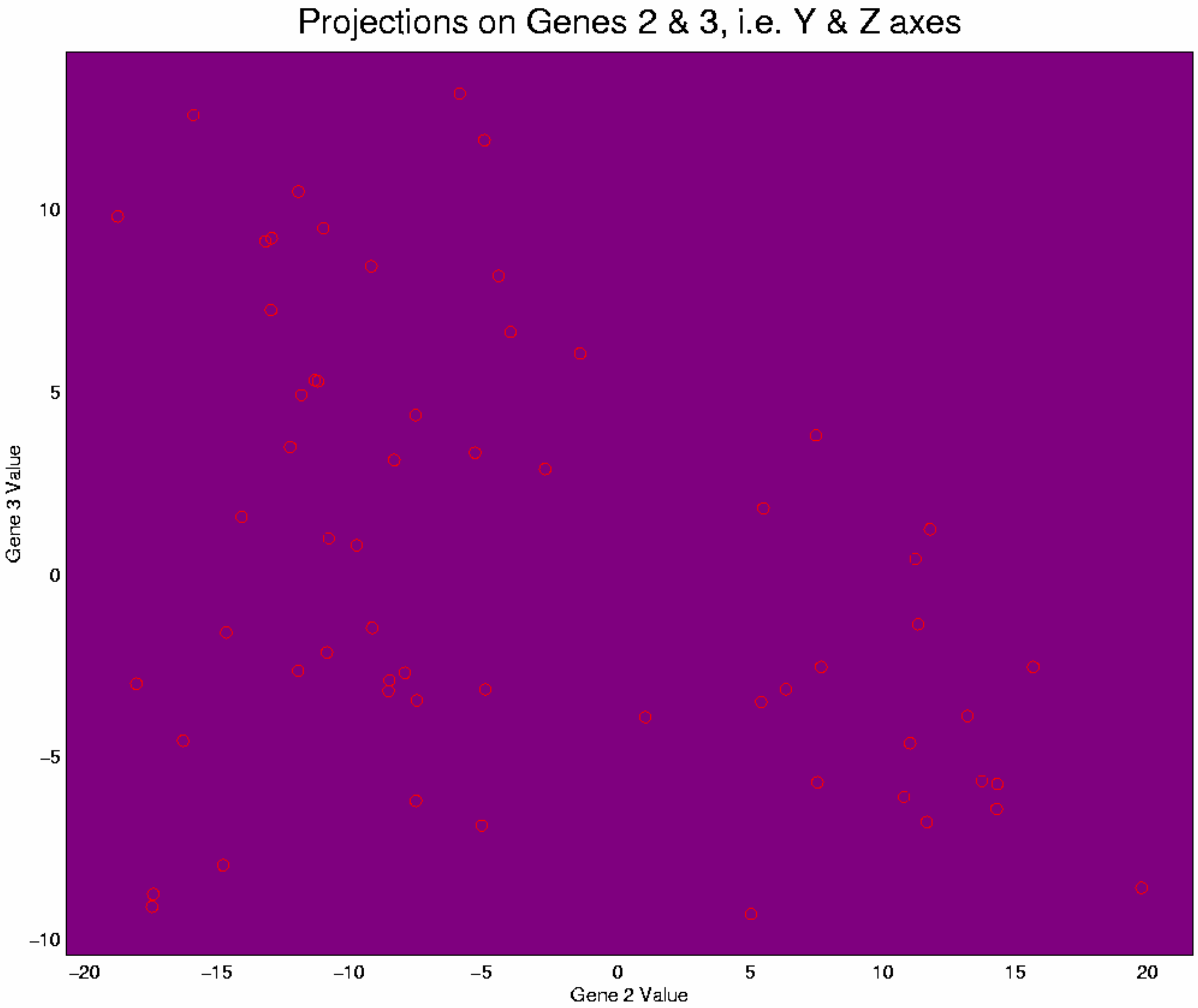


Illust’n of Multivar. View: 2-d Proj’n, YZ-plane



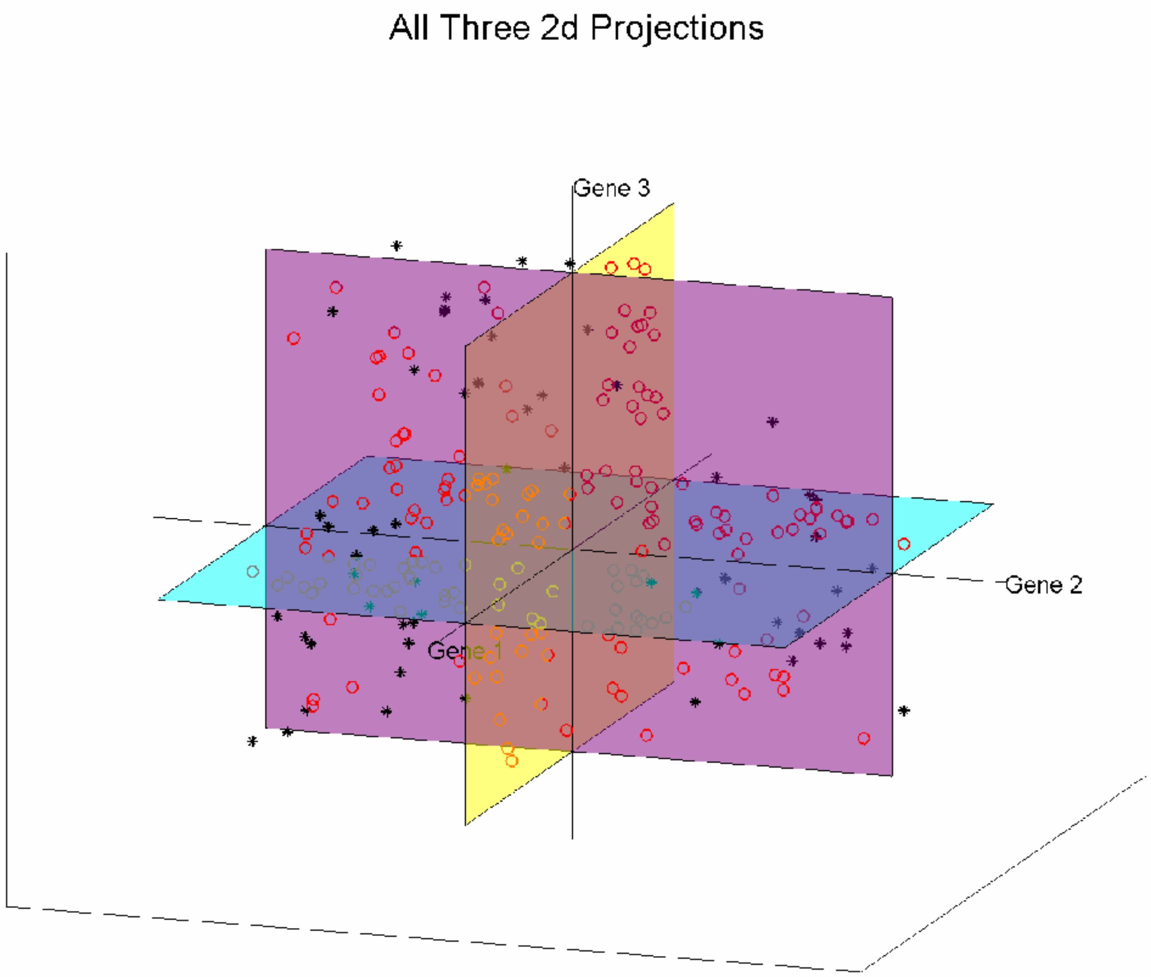
UNC Lineberger

Illust’n of Multivar. View: YZ-Proj’n, 2-d view



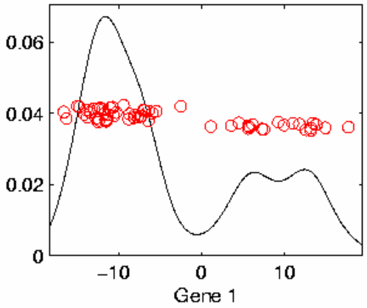
Illust'n of Multivar. View: all 3 planes

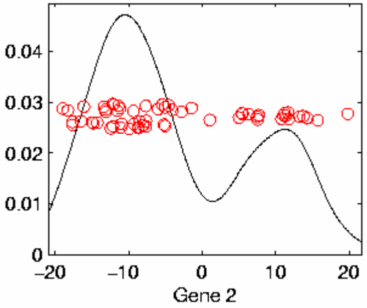
UNC Lineberger

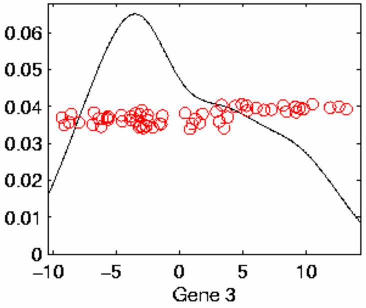


UNC Lineberger

Illust'n of Multivar. View: Diagonal 1-d proj'ns

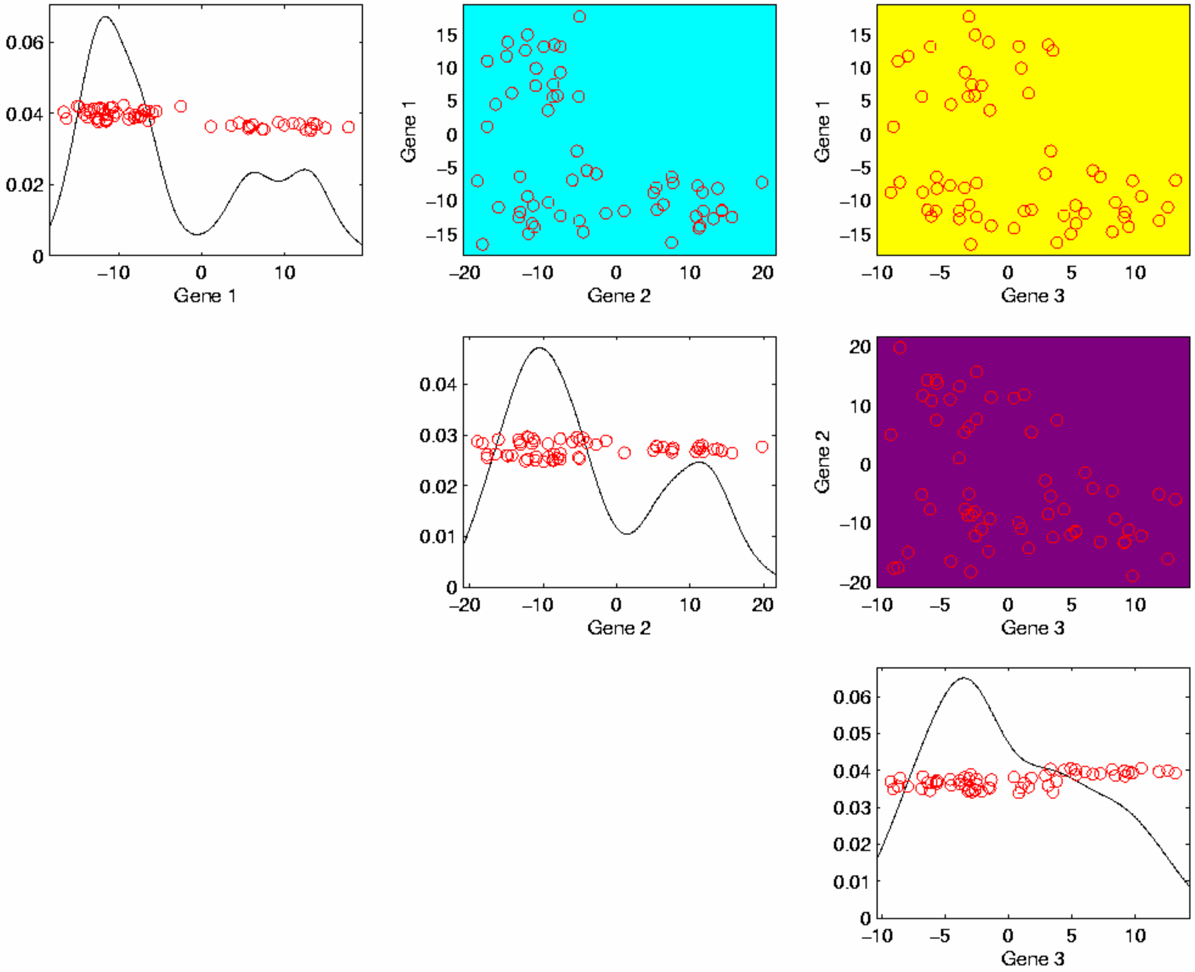






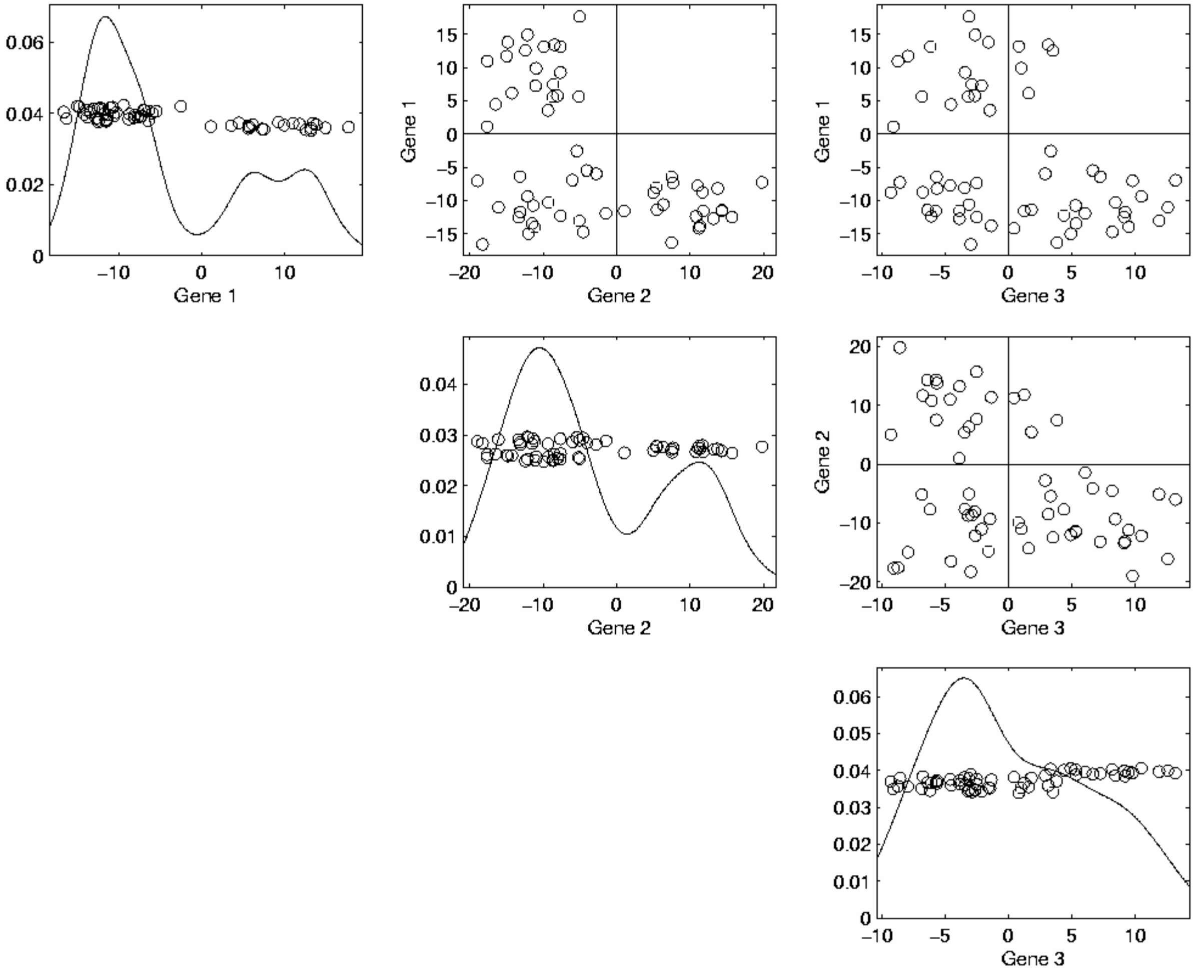
UNC Lineberger

Illust’n of Multivar. View: Add off-diagonals



UNC Lineberger

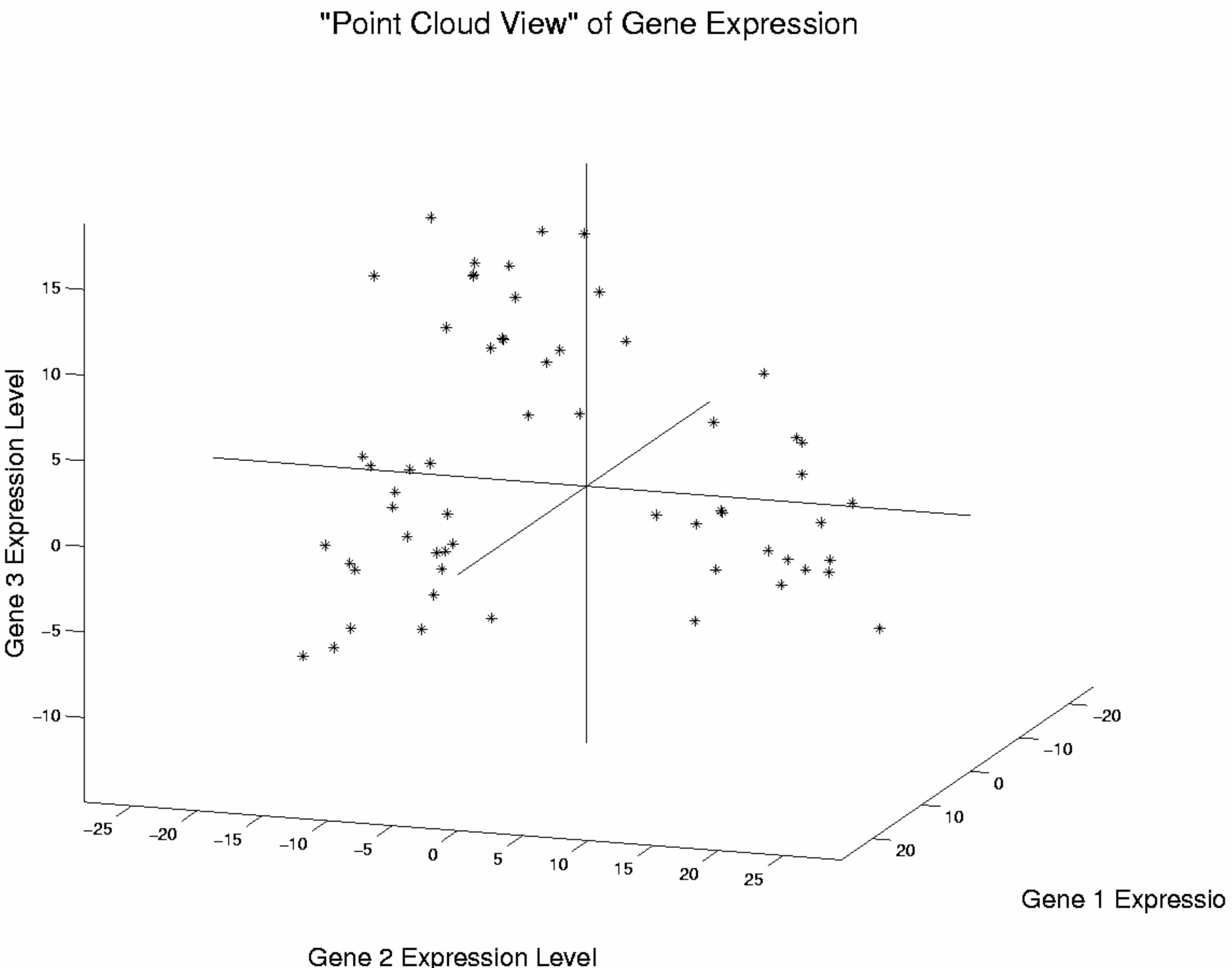
Illust'n of Multivar. View: Typical View



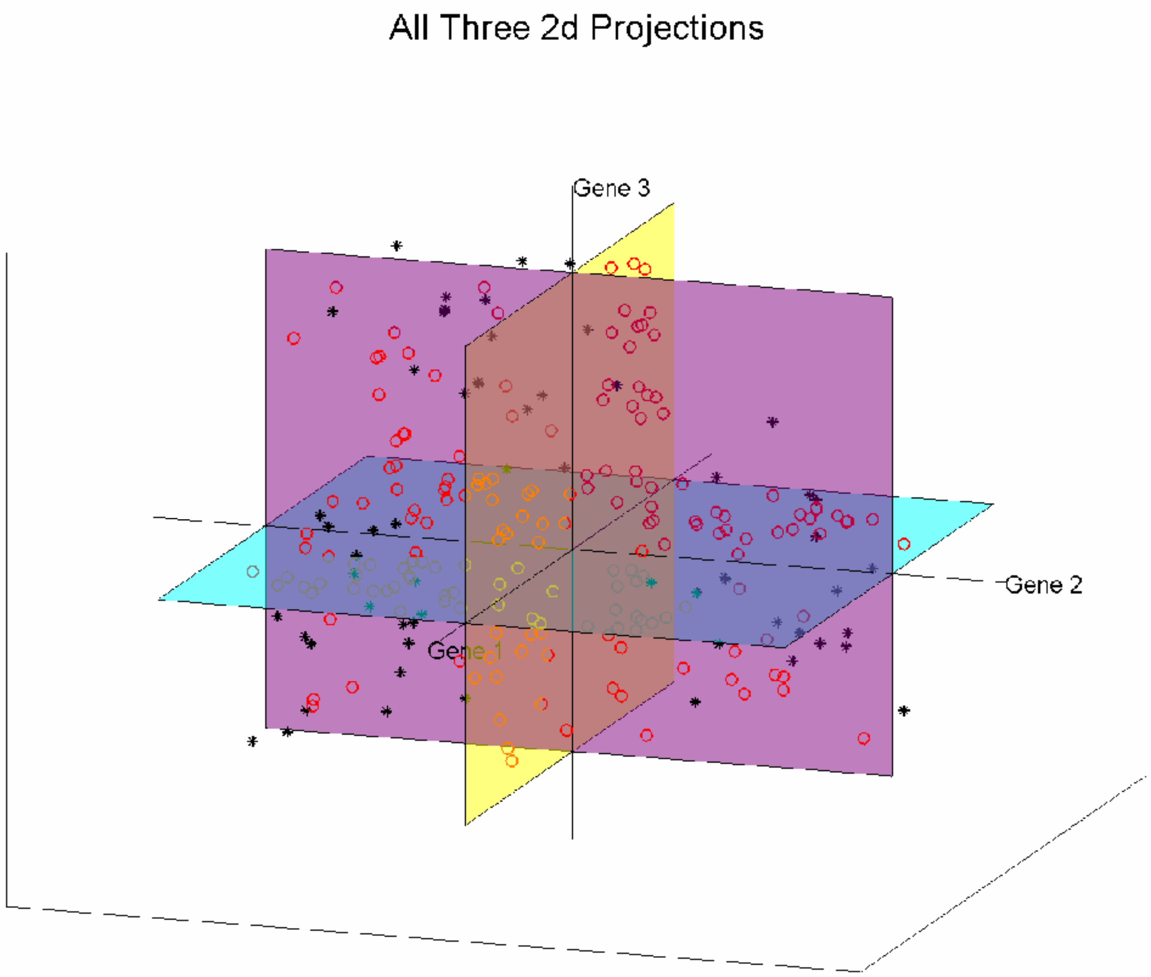
Improved View

- Idea: “rotations” of coordinate systems
  - More useful views
- Generally consider “useful directions”
- E.g. 1: Principal Component directions
  - Directions that “maximize variation”
  - Often insightful
  - Also called “eigengenes” or “metagenes”
- E.g. 2: DWD directions
  - Directions that “maximize separation”
  - DWD = “Distance Weighted Discrimination”
  - Improved version of SVM

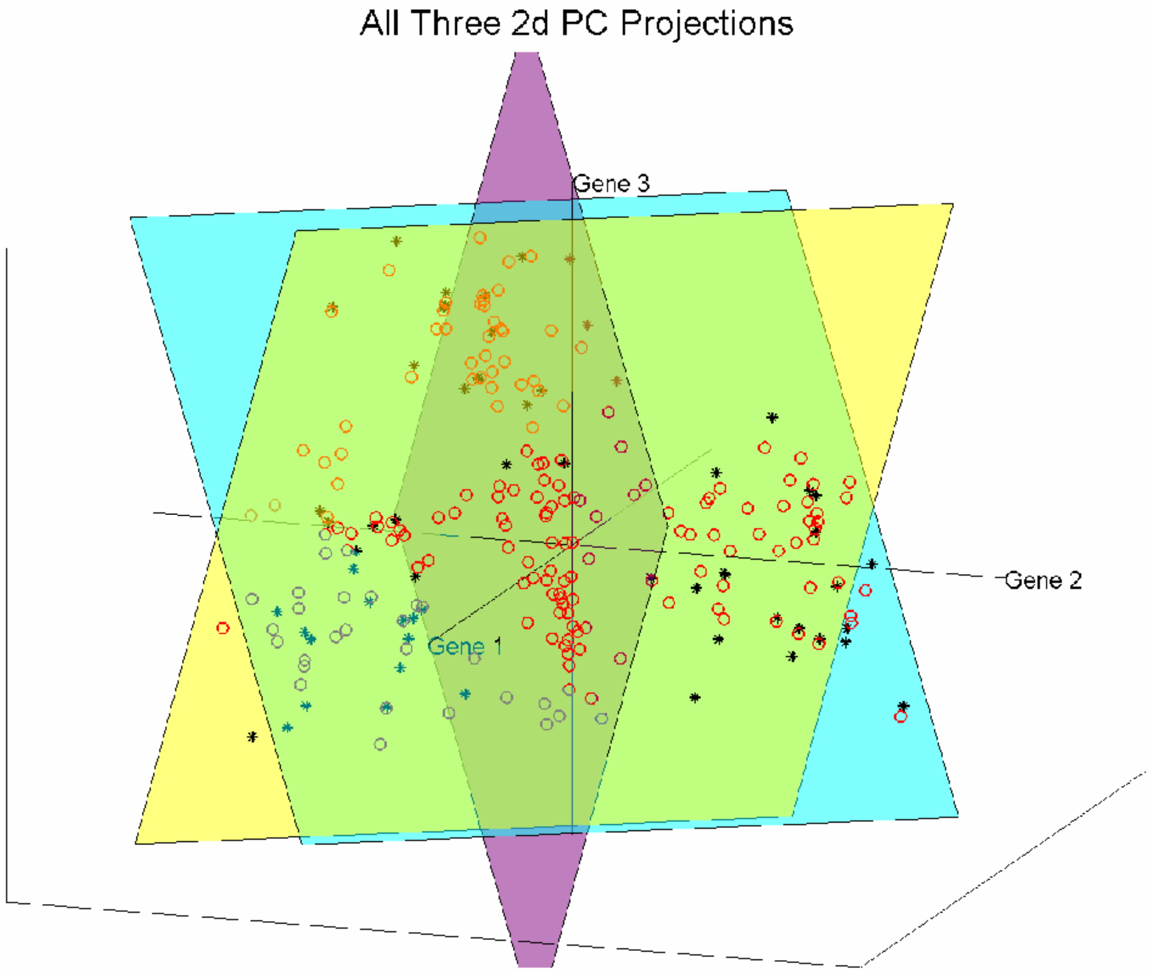
Illust'n of PCA View: Recall Raw Data



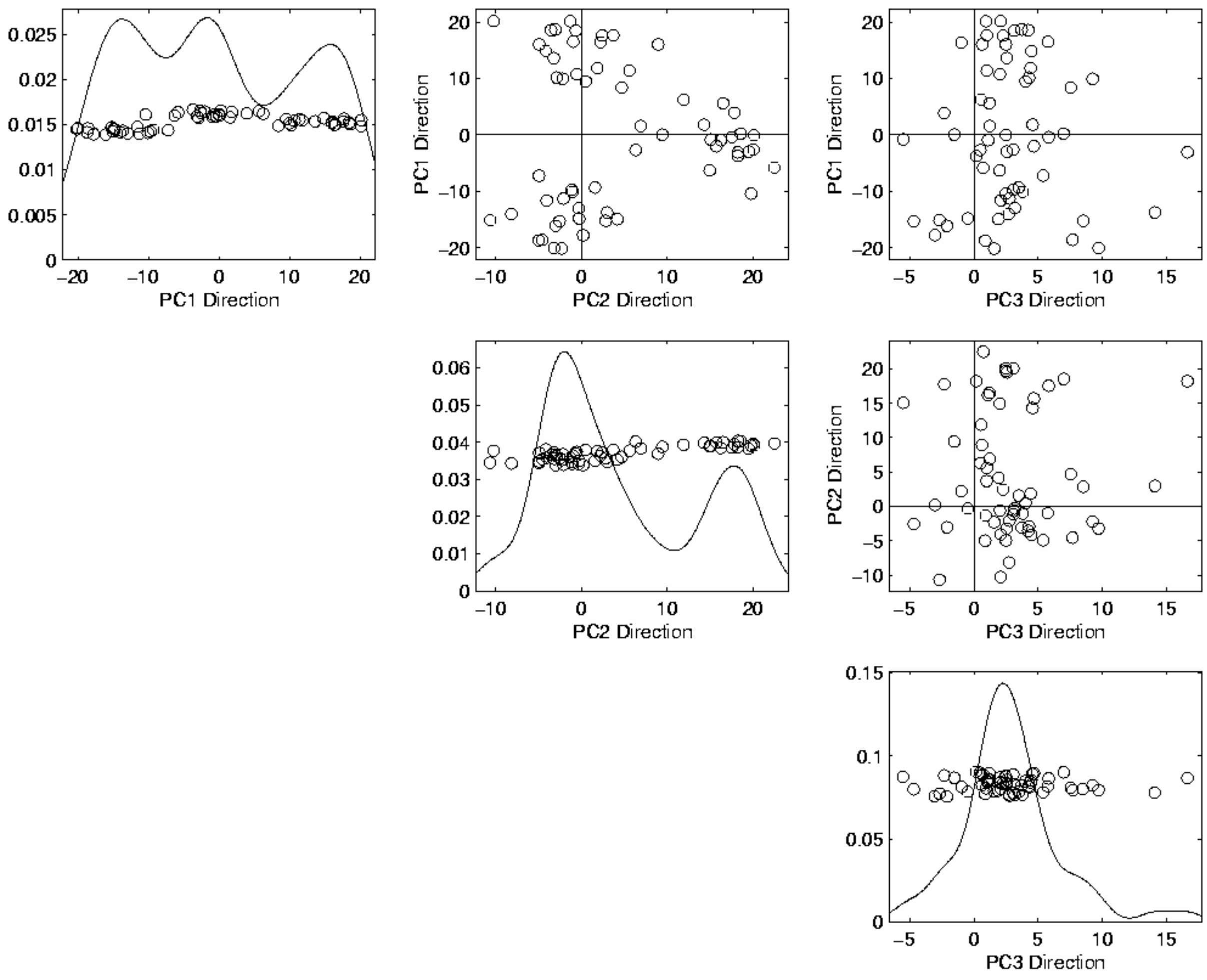
Illust'n of PCA View: Recall Gene by Gene Views



Illust'n of PCA View: All 3 PC Projections



Illust’n of PCA View: Typical View

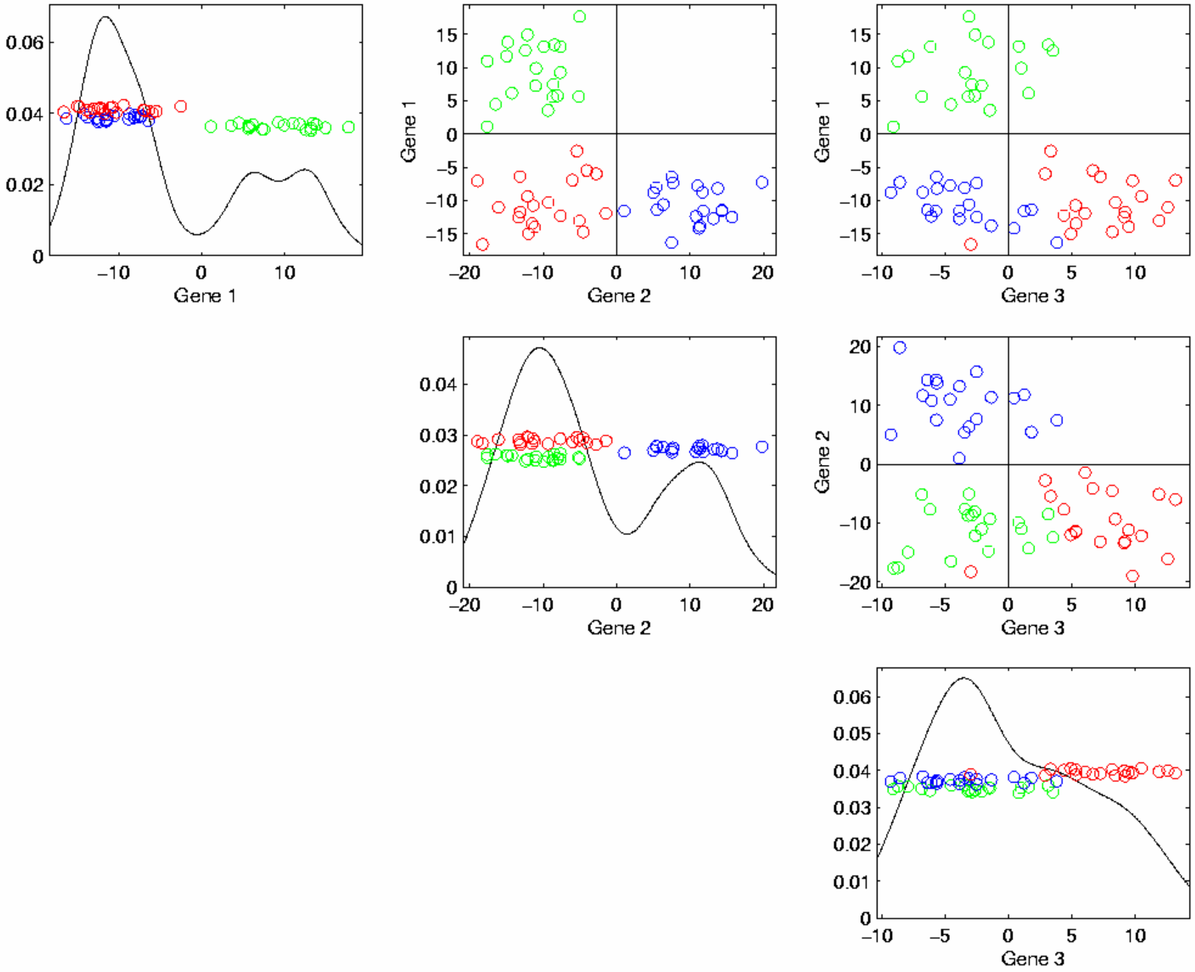


Comparison of Views

- Highlight 3 clusters
- Gene by Gene View
  - Clusters appear in all 3 scatterplots
  - But never very separated
- PCA View
  - 1st shows three distinct clusters
  - Better separated than in gene view
  - Clustering concentrated in 1st scatterplot
- Effect is small, since only 3-d

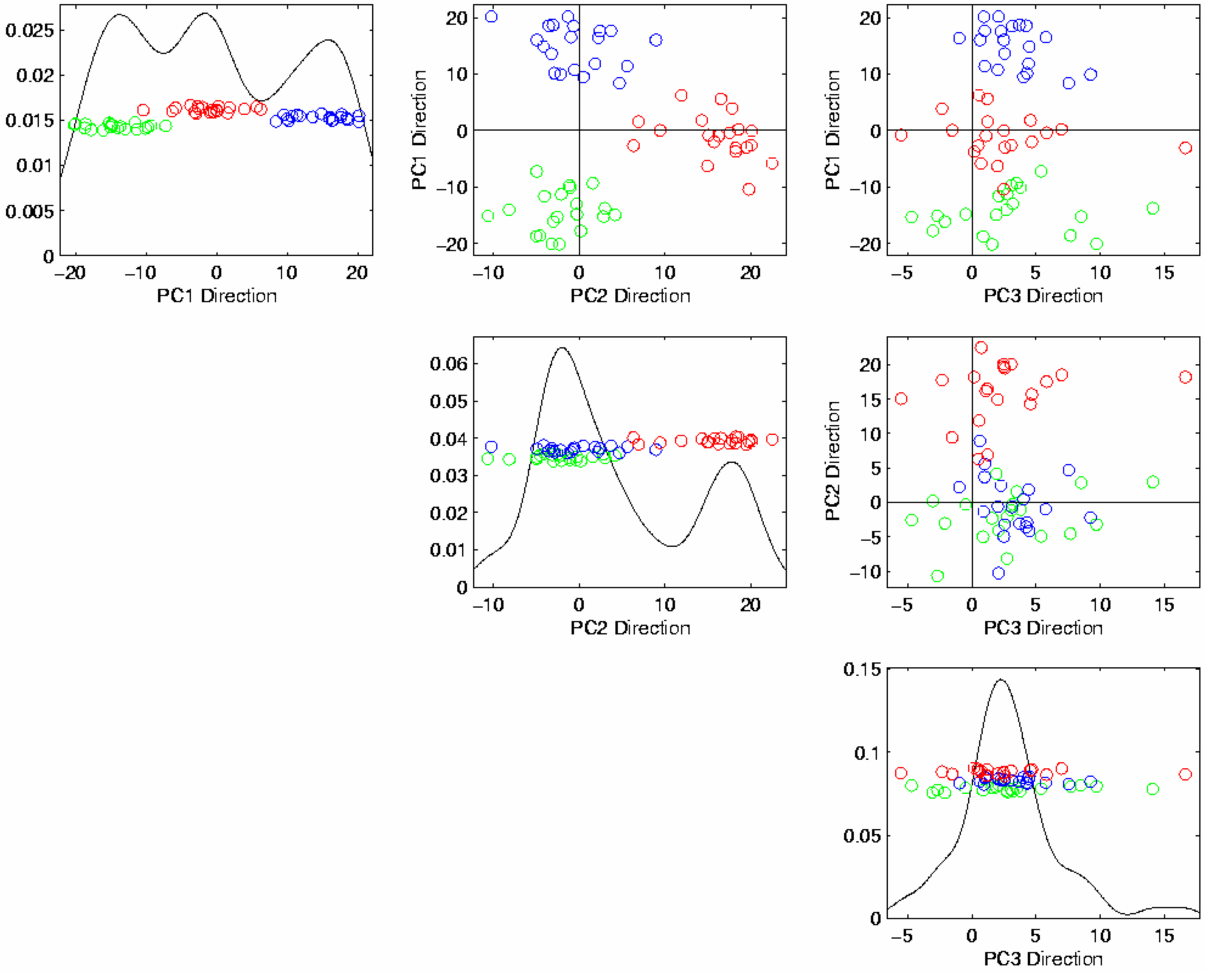
Illust’n of PCA View: Gene by Gene View

UNC Lineberger



UNC Lineberger

Illust'n of PCA View: PCA View

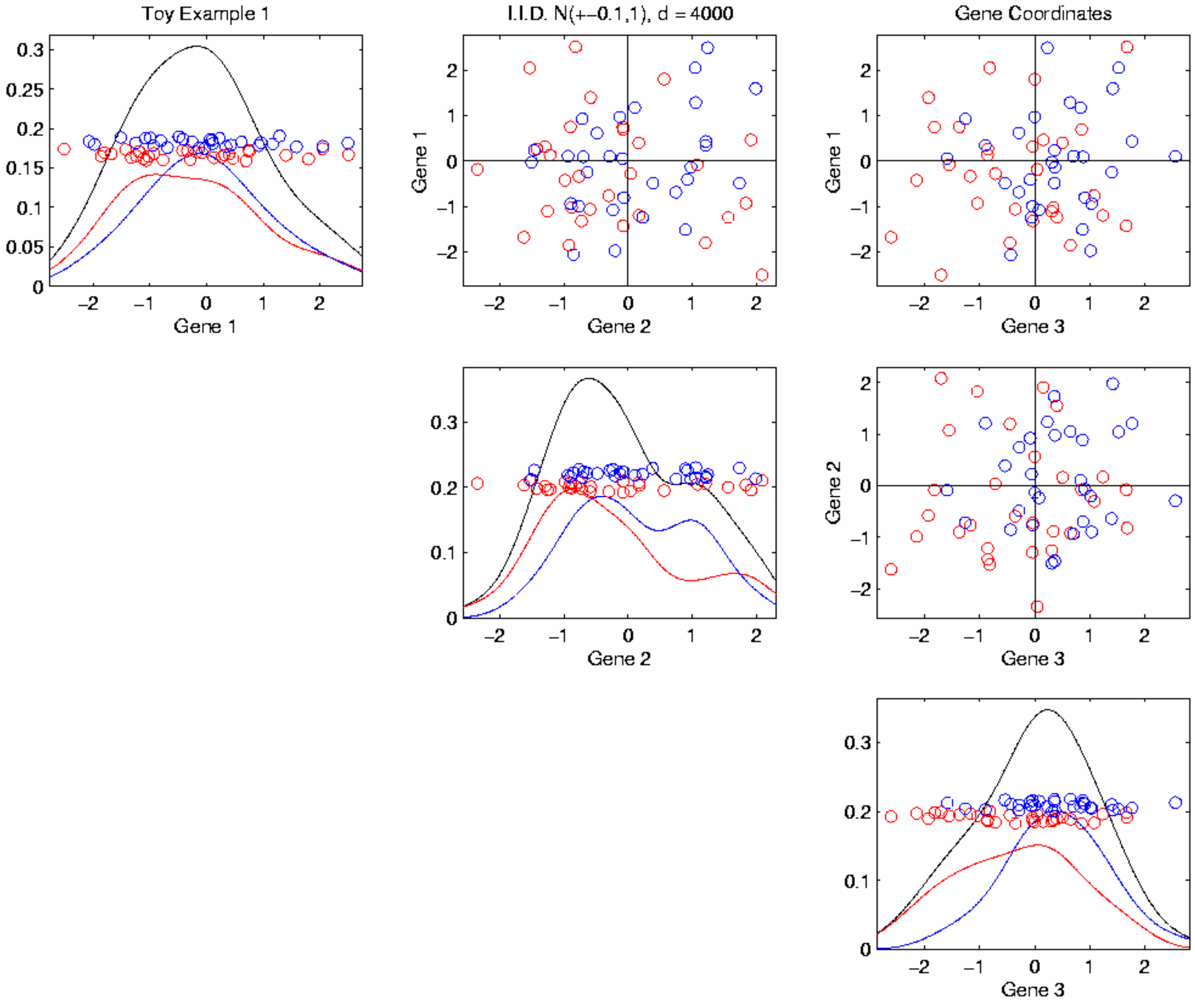


Another Comparison of Views

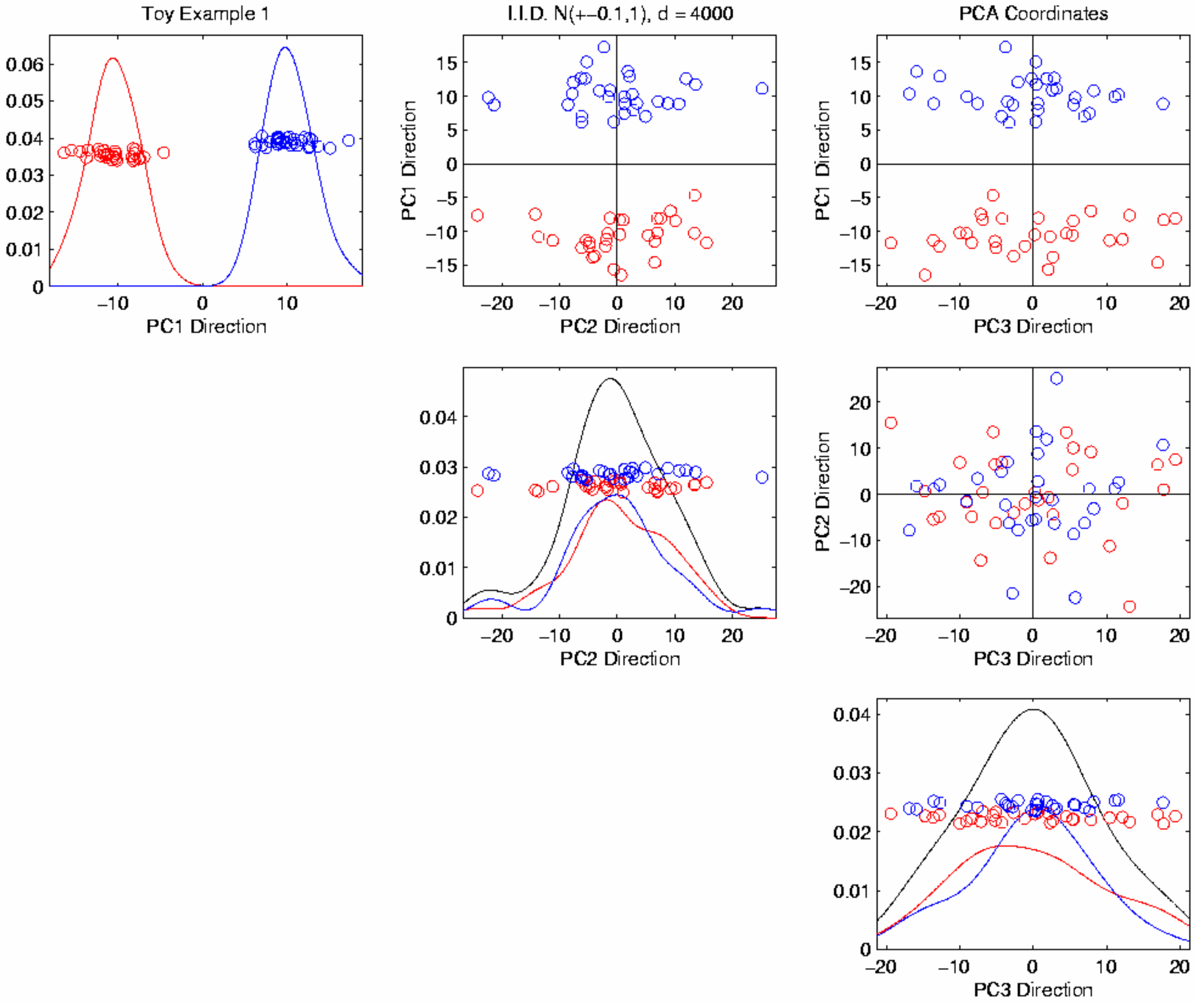
- Much higher dimension, # genes = 4000
- Gene by Gene View
  - Clusters very nearly the same
  - Very slight difference in means
- PCA View
  - Huge difference in 1st PC Direction
  - Magnification of clustering
  - Lesson: Alternate views can show much more (especially in high dimensions, i.e. for many genes)
  - Shows PC view is very useful

Another Comparison: Gene by Gene View

UNC Lineberger



Another Comparison: PCA View



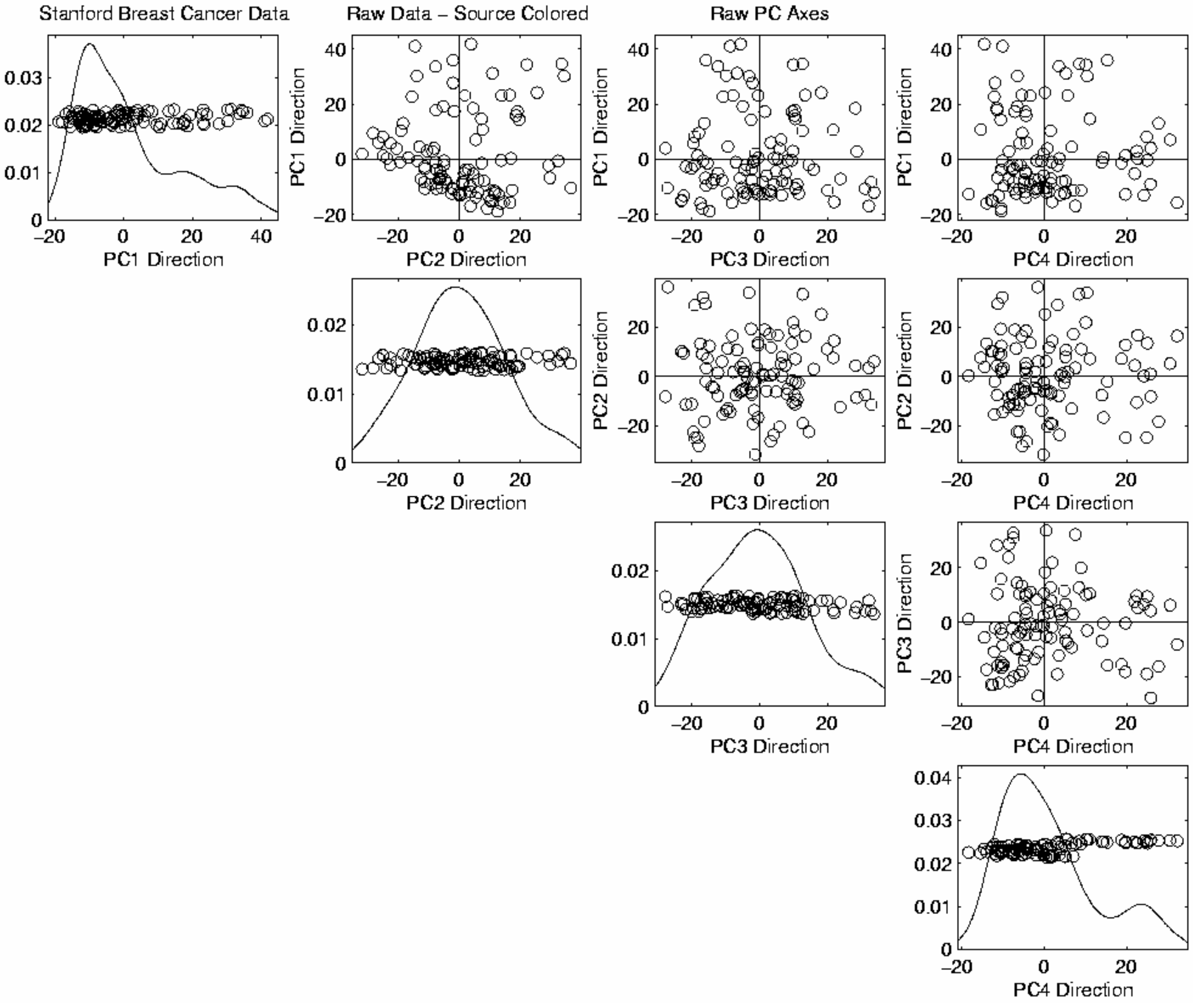
Batch and Source Adjustment

- For Stanford Breast Cancer Data
- Analysis in Benito, et al (2004) *Bioinformatics*

  https://genome.unc.edu/pubsup/dwd/
- Adjust for Source Effects
  - Different sources of mRNA
- Adjust for Batch Effects
  - Arrays fabricated at different times
  - Batches were shared between labs

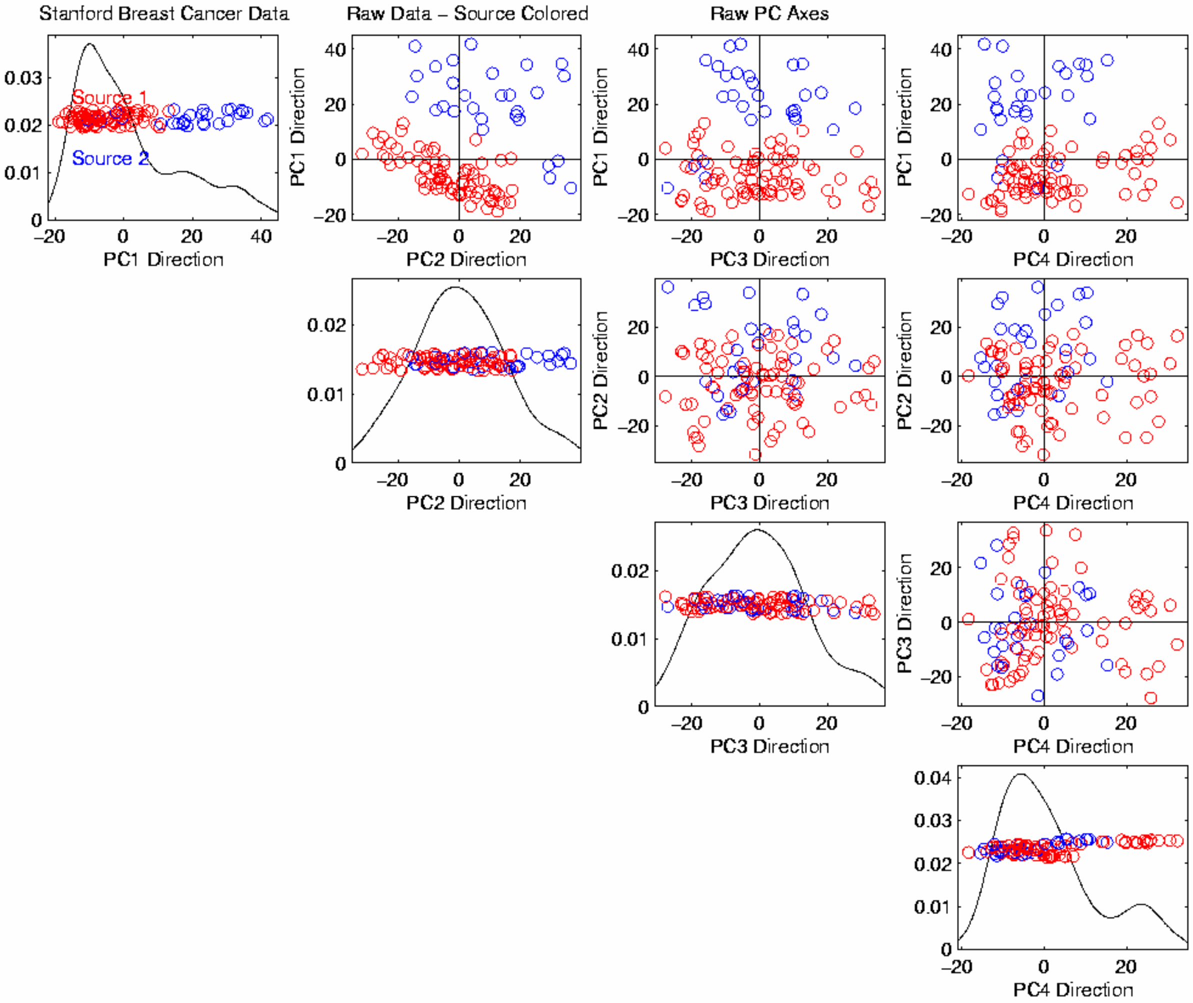
UNC Lineberger

Source Batch Adj: Raw Breast Cancer data



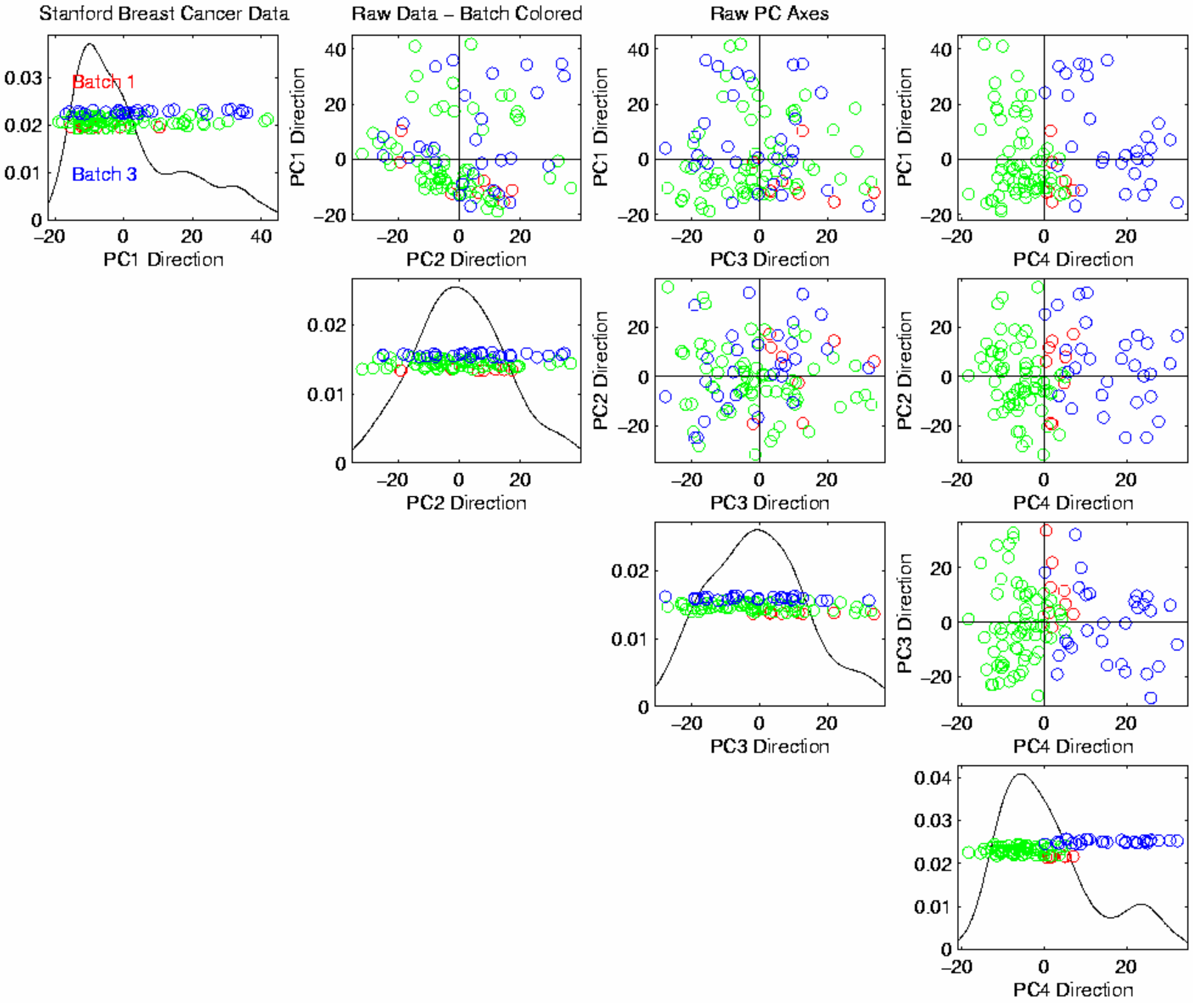
UNC Lineberger

Source Batch Adj: Source Colors

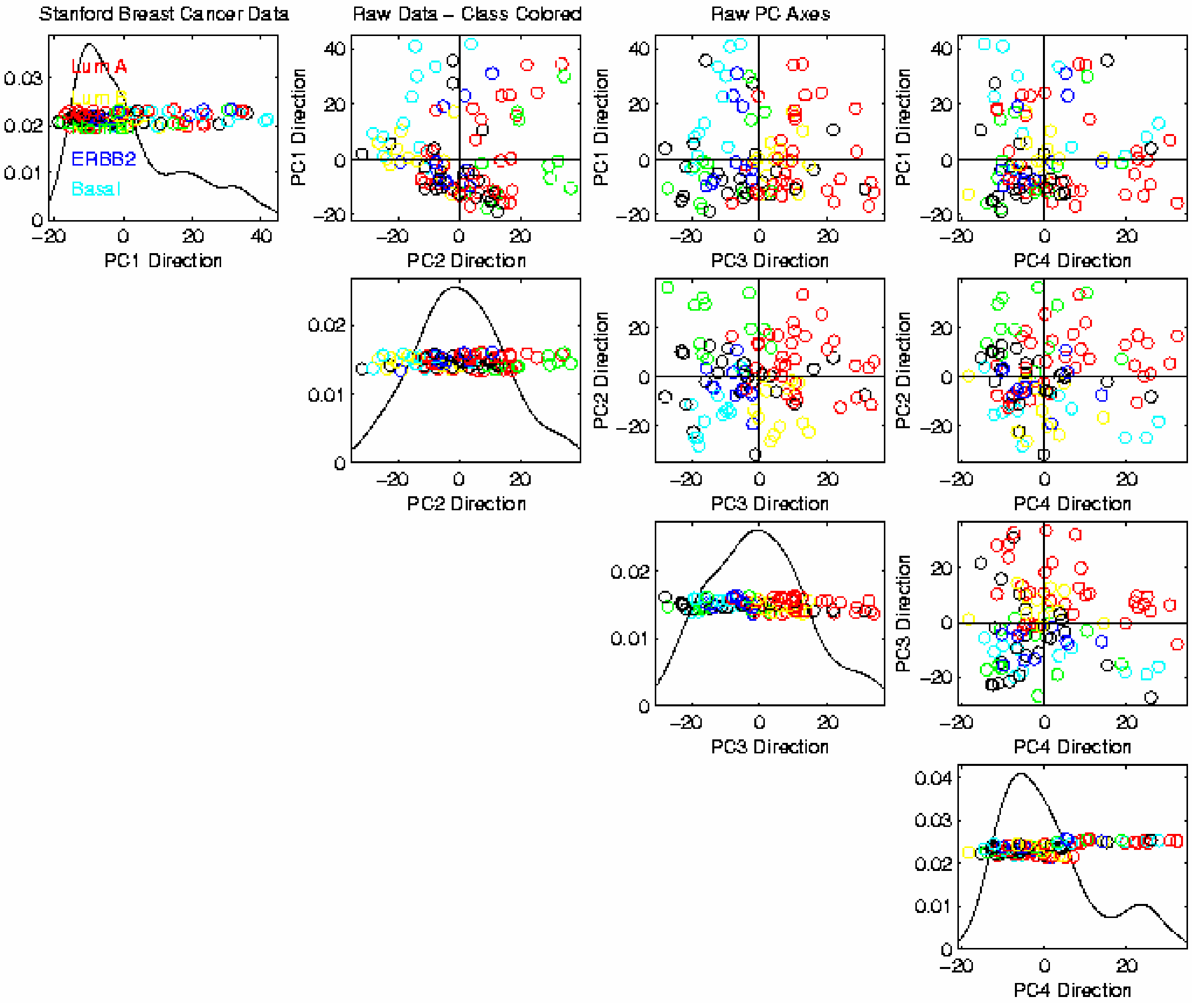


UNC Lineberger

Source Batch Adj: Batch Colors

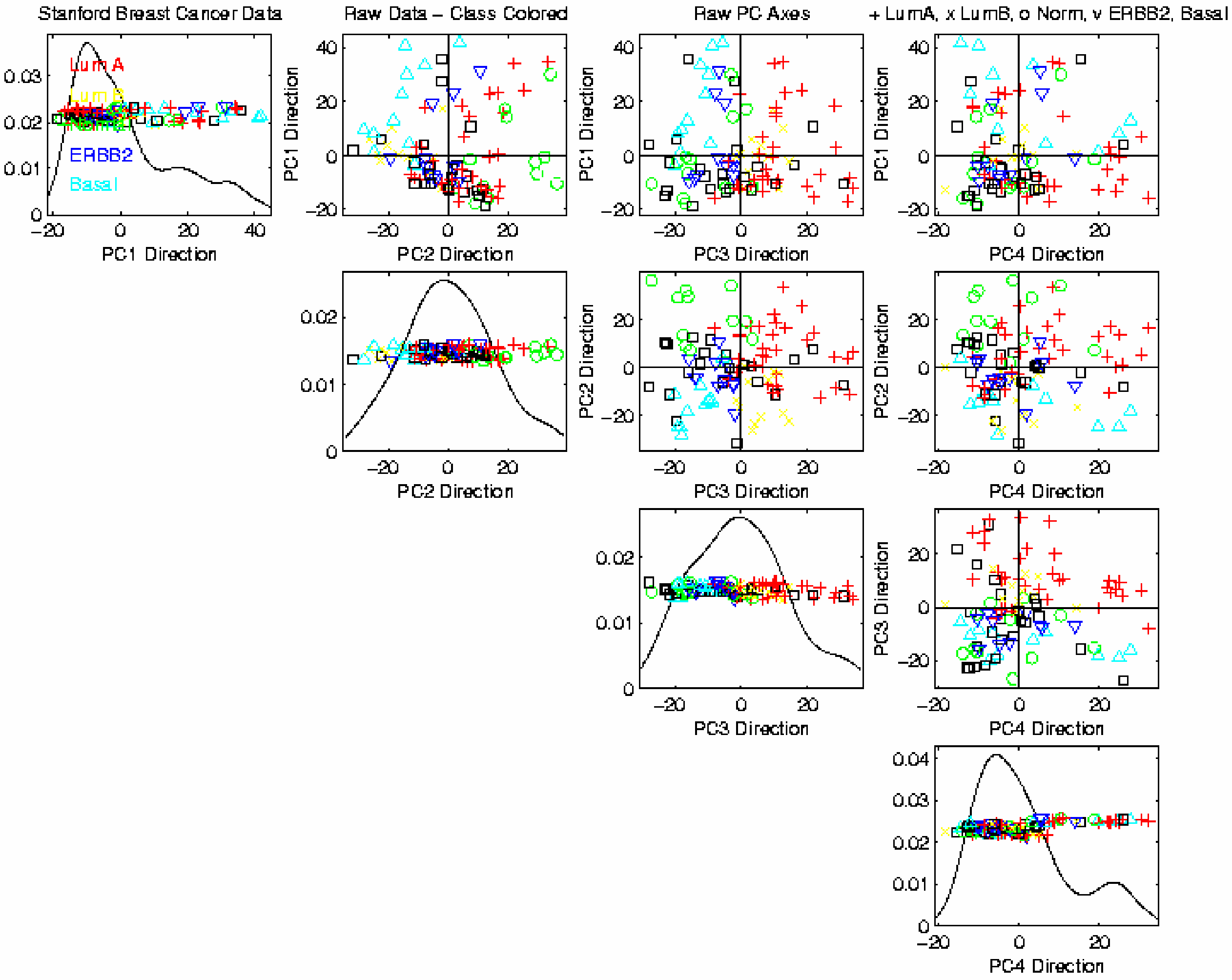


Source Batch Adj: Biological Class Colors



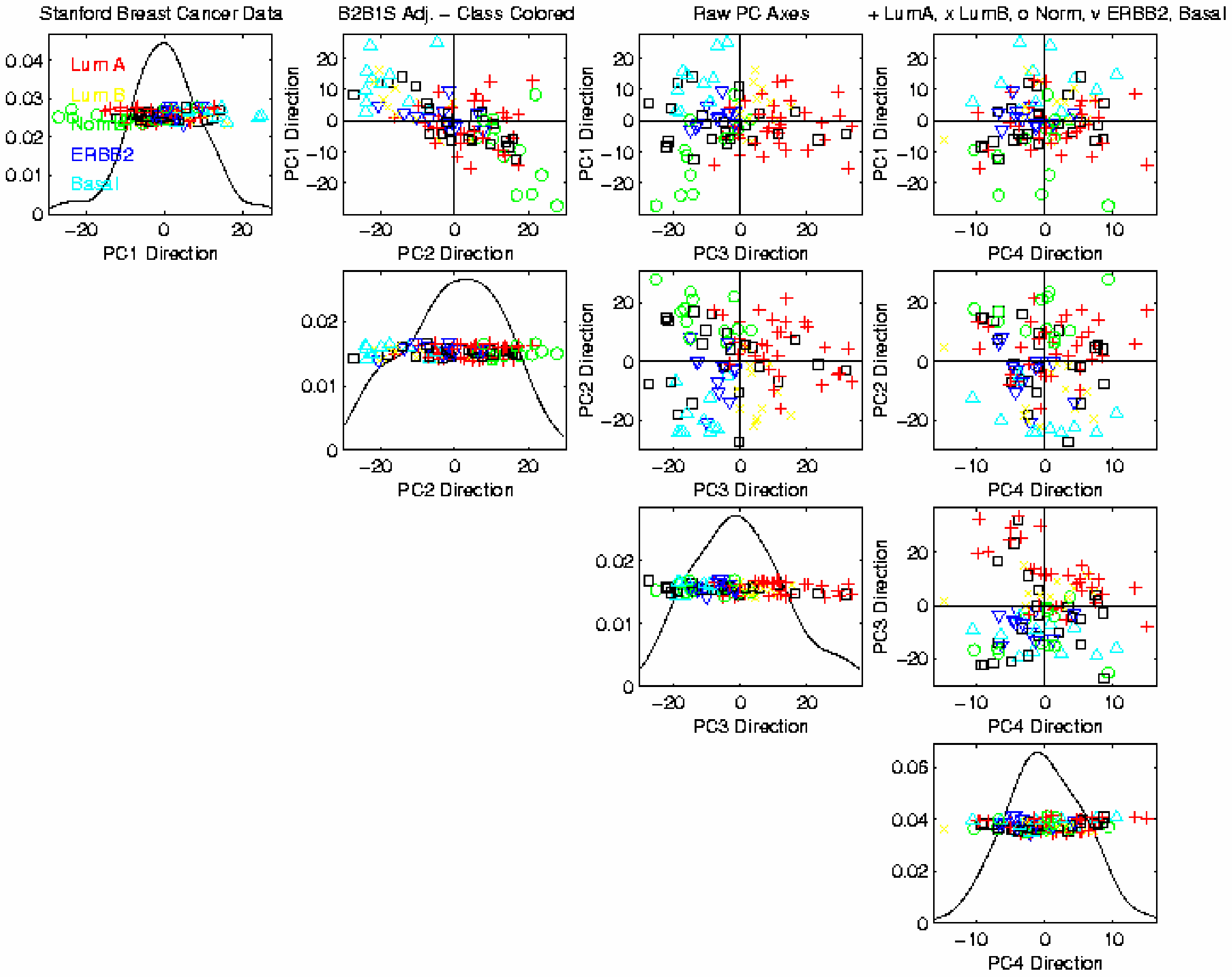
UNC Lineberger

Source Batch Adj: Biological Class Col. & Symbols



UNC Lineberger

Source Batch Adj: S. & B Adj'd, Class Colors



Source Batch Adj: More Views

Internet Available:

http://genome.med.unc.edu:8080/caBIG/DWDindex.htm

Follow Link:

DWD Bias Adjustment of Batch and Source Effects

Interesting Benchmark Data Set

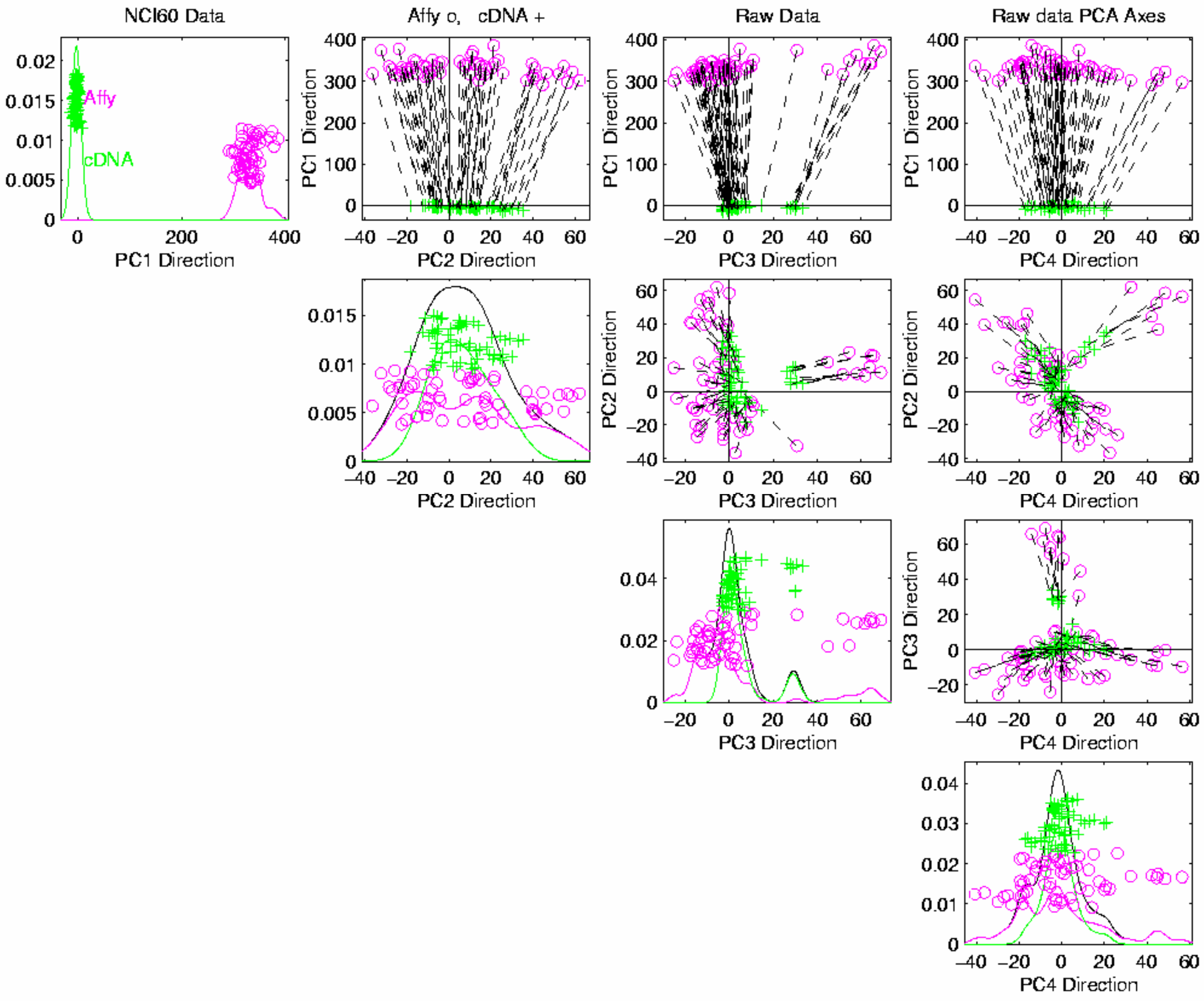
- NCI 60 Cell Lines
  - Interesting benchmark, since *same* cells
  - Data Web available:
    http://discover.nci.nih.gov/datasetsNature2000.jsp
  - *Both* cDNA and Affymetrix Platforms

- Different from Breast Cancer Data
  - No common RNA

- Interest in “mixed samples”???

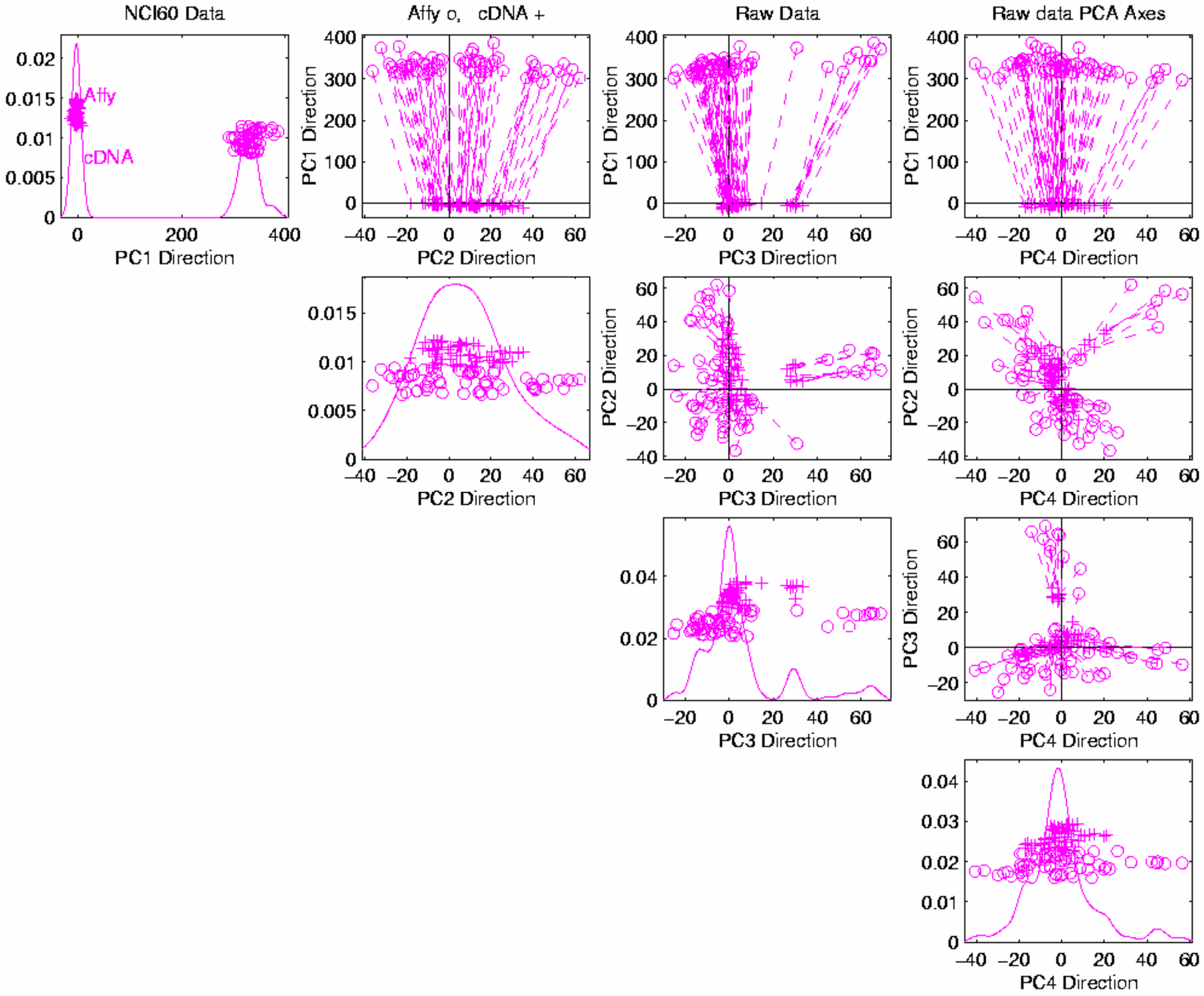
UNC Lineberger

NCI 60: Raw Data, Platform Colored



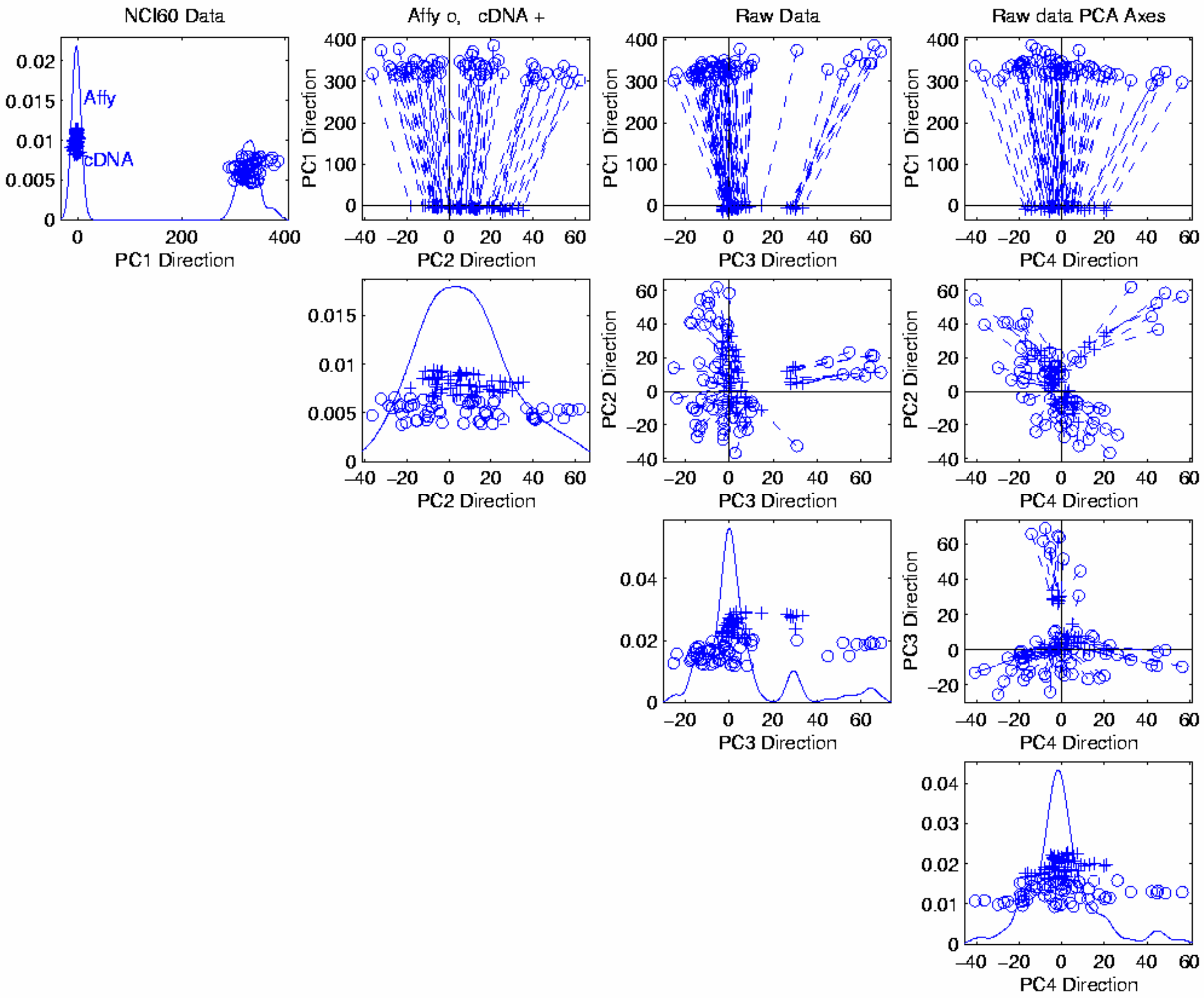
UNC Lineberger

NCI 60: Raw Data



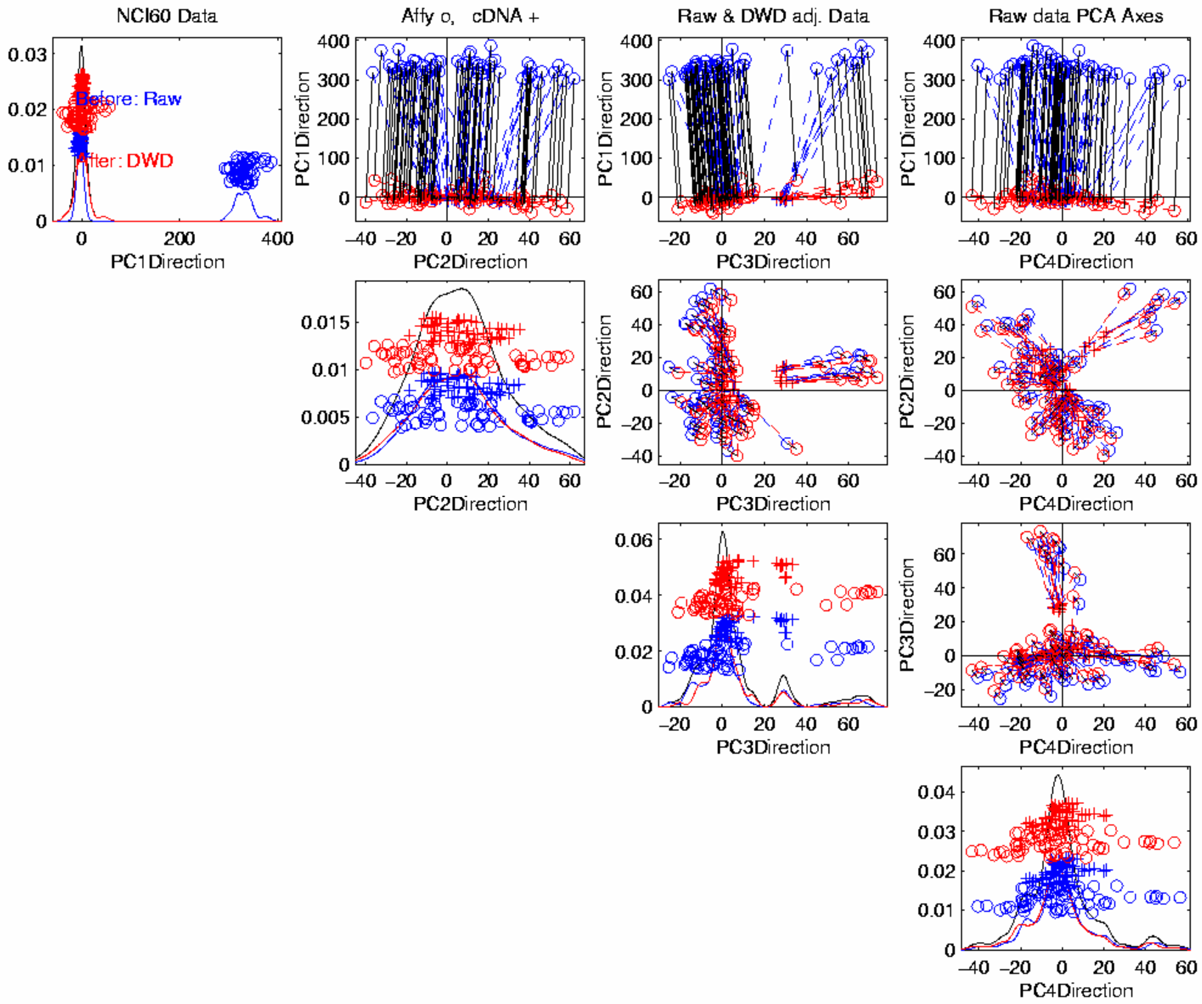
UNC Lineberger

NCI 60: Raw Data, Before DWD Adjustment



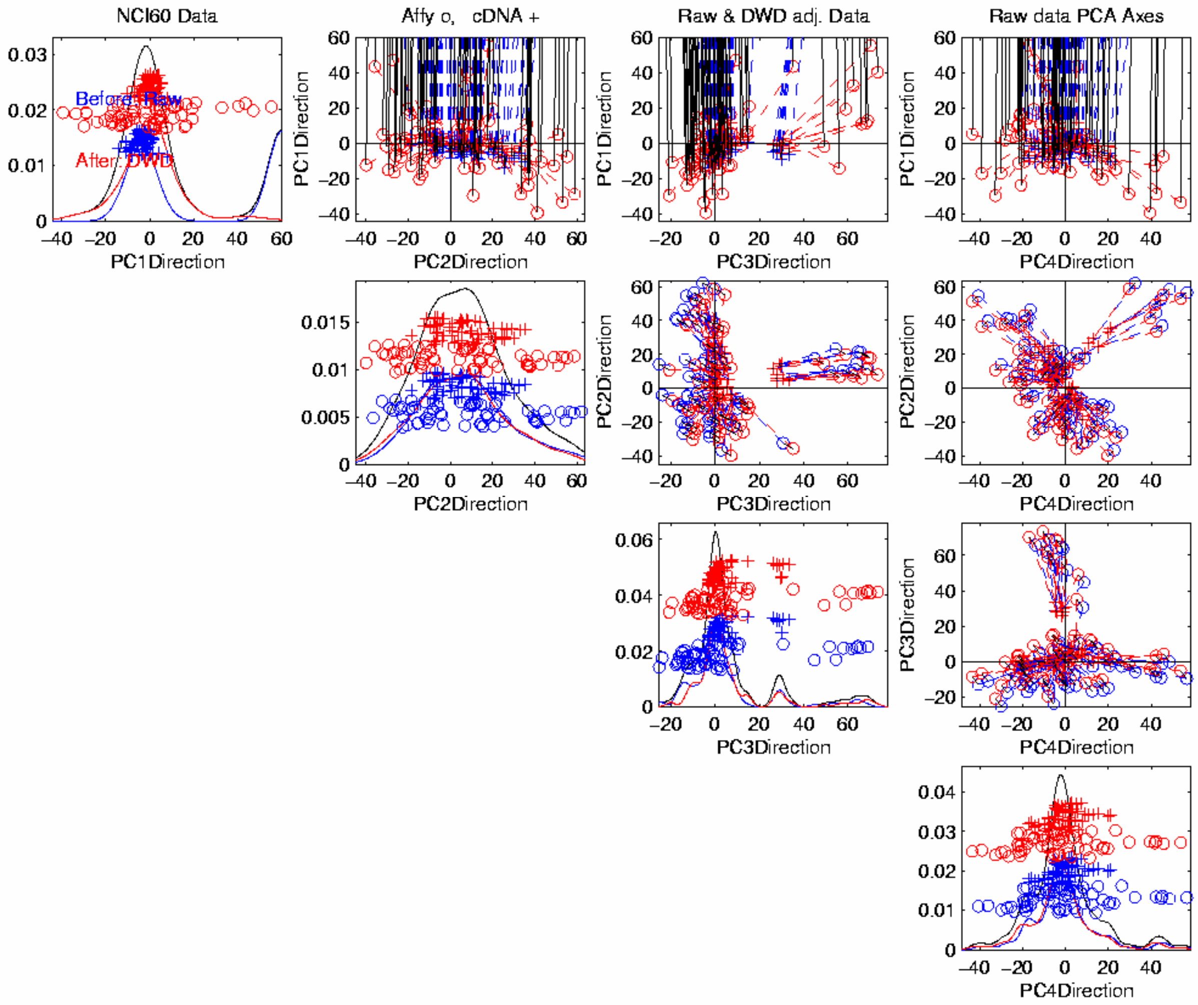
UNC Lineberger

NCI 60: Before & After DWD adjustment



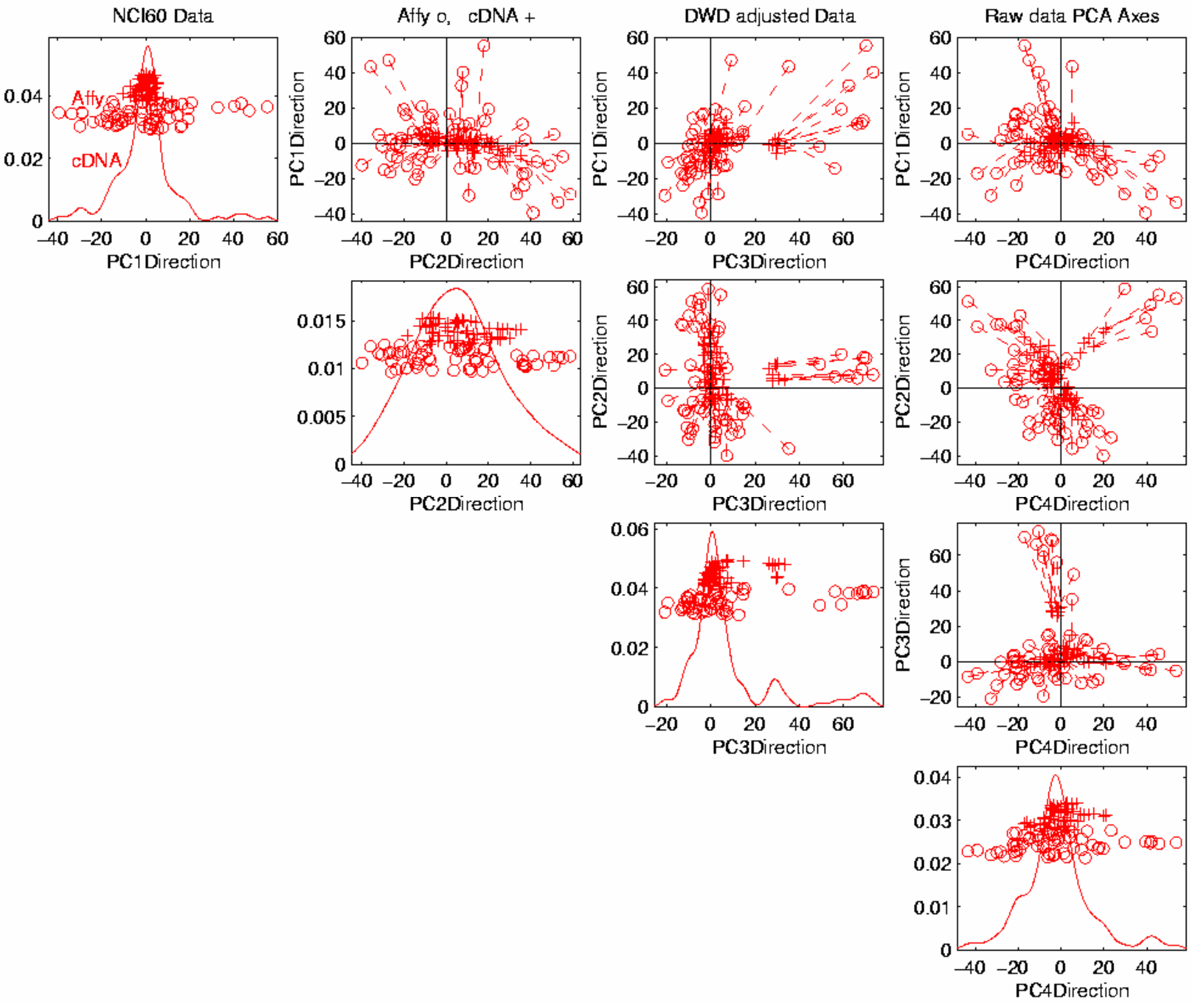
UNC Lineberger

NCI 60: Before & After, new scales

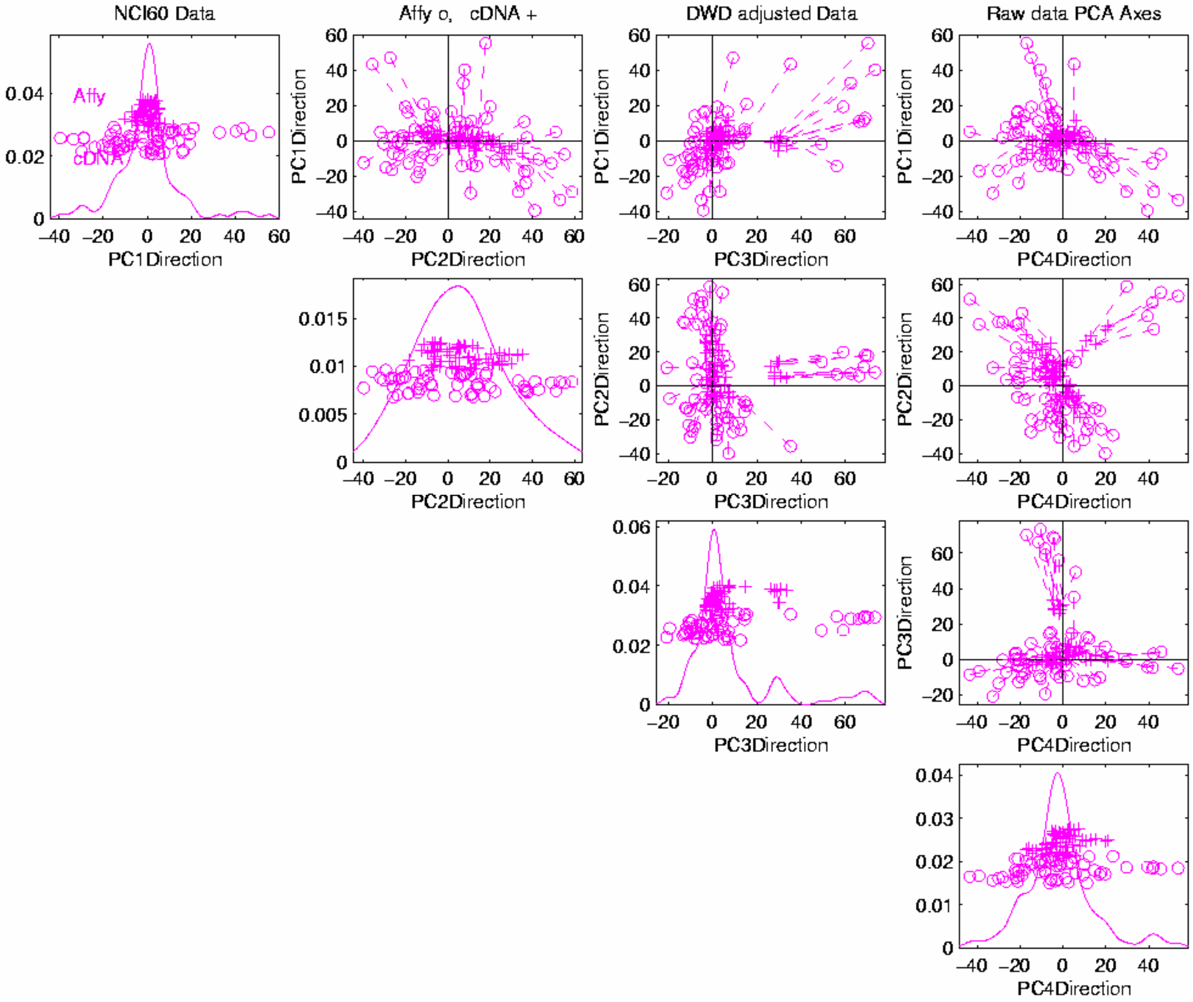


UNC Lineberger

NCI 60: After DWD

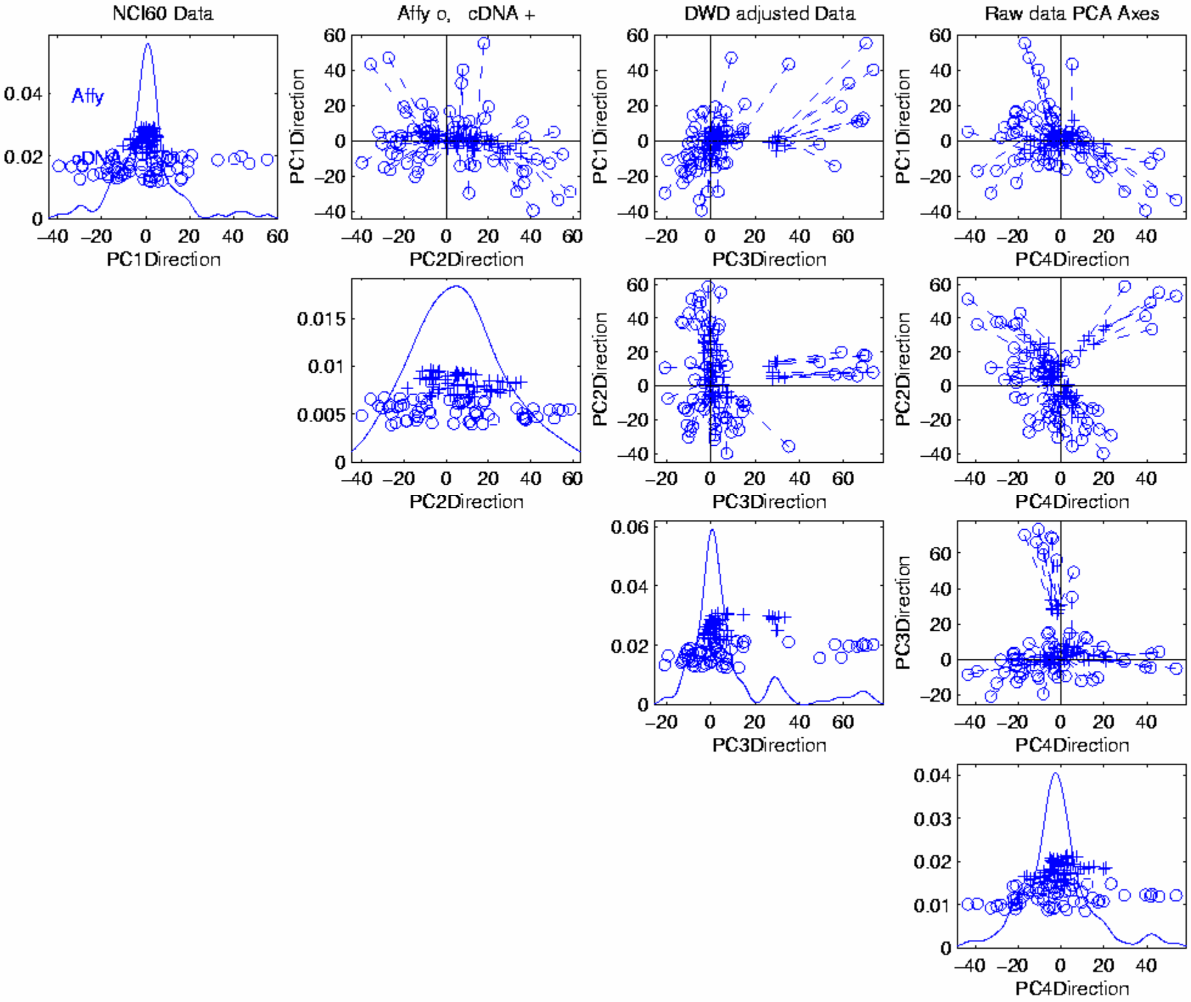


NCI 60: DWD adjusted data

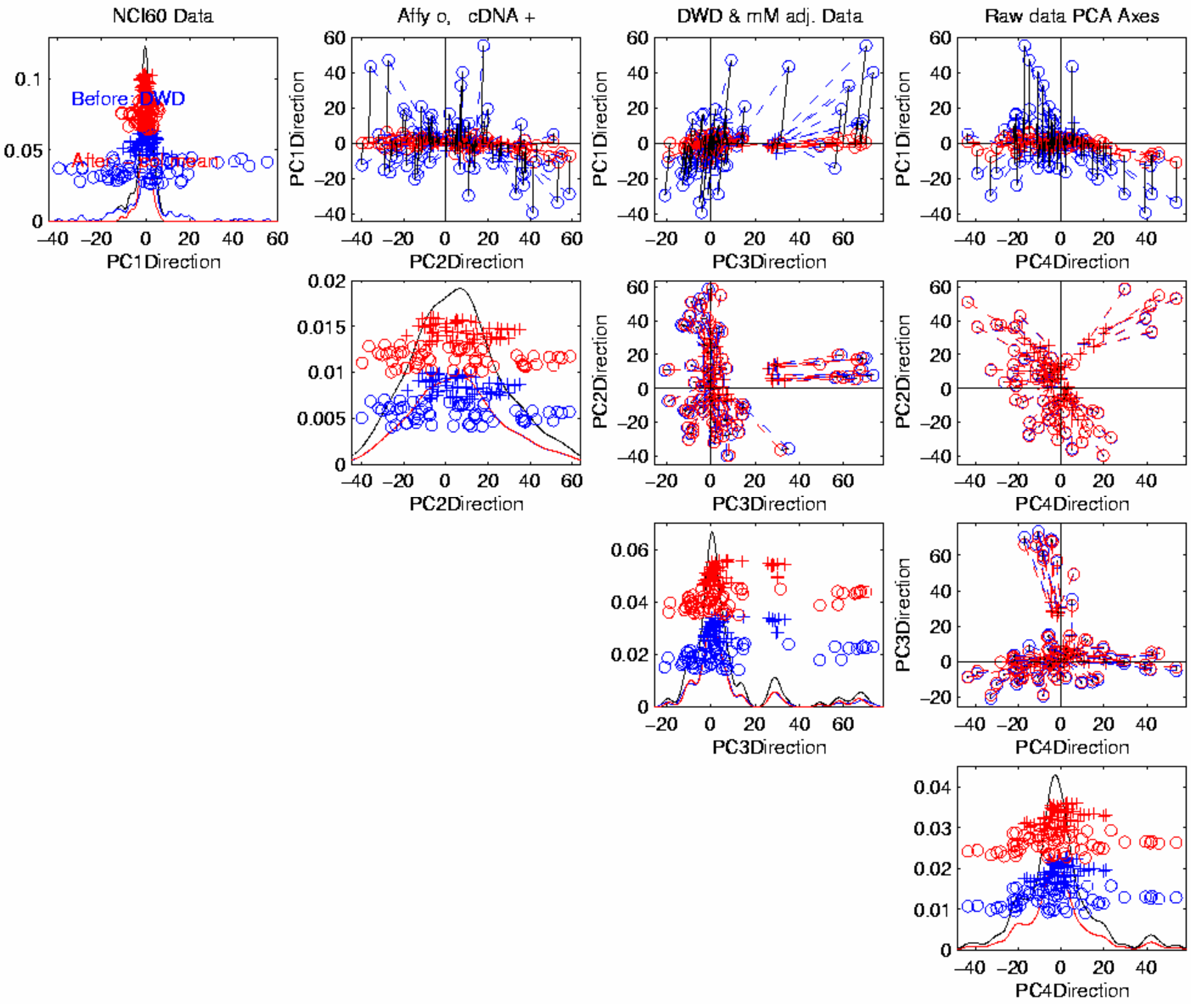


UNC Lineberger

NCI 60: Before Column Mean Adjustment

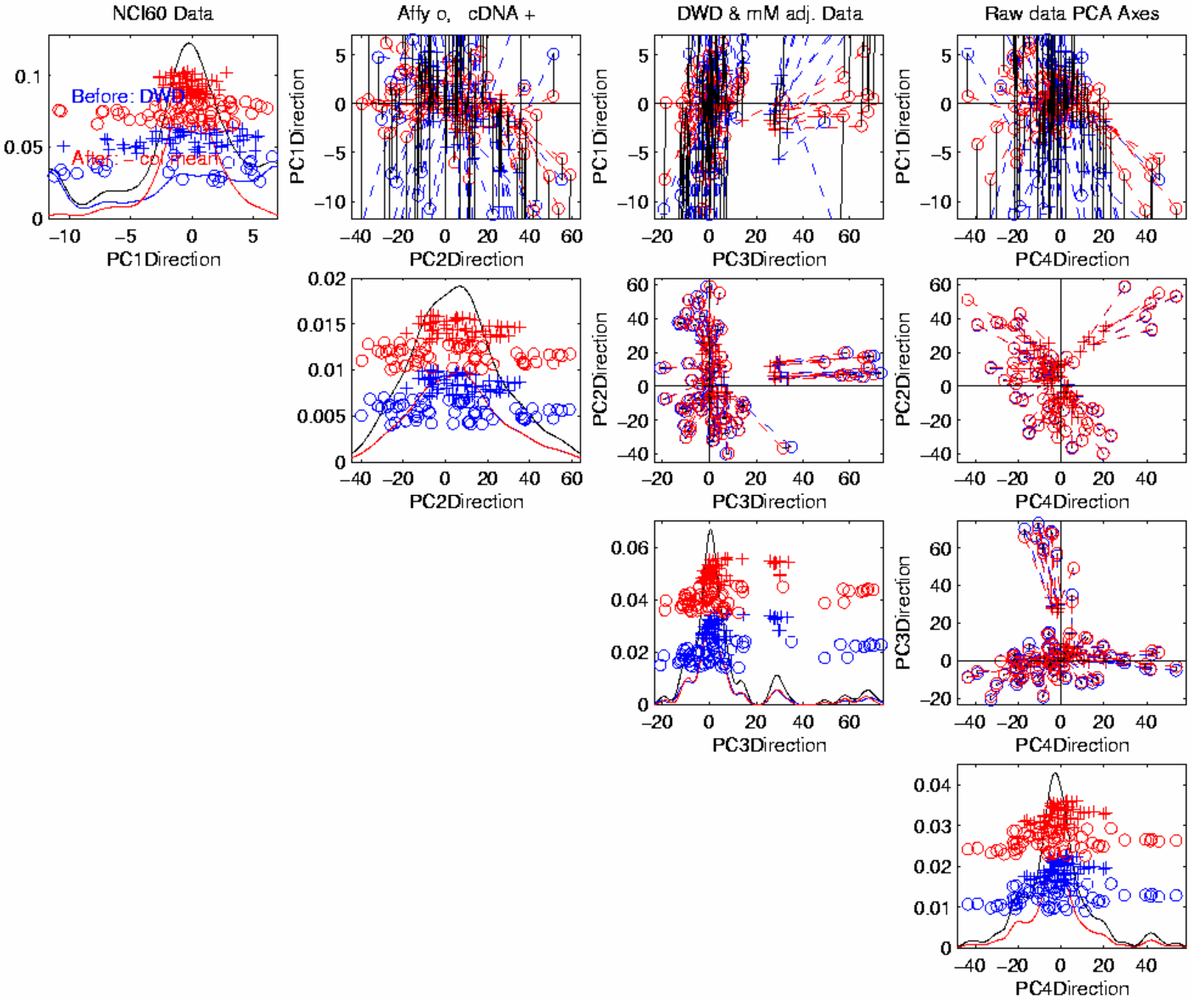


NCI 60: Before & After Column Mean Adjustment



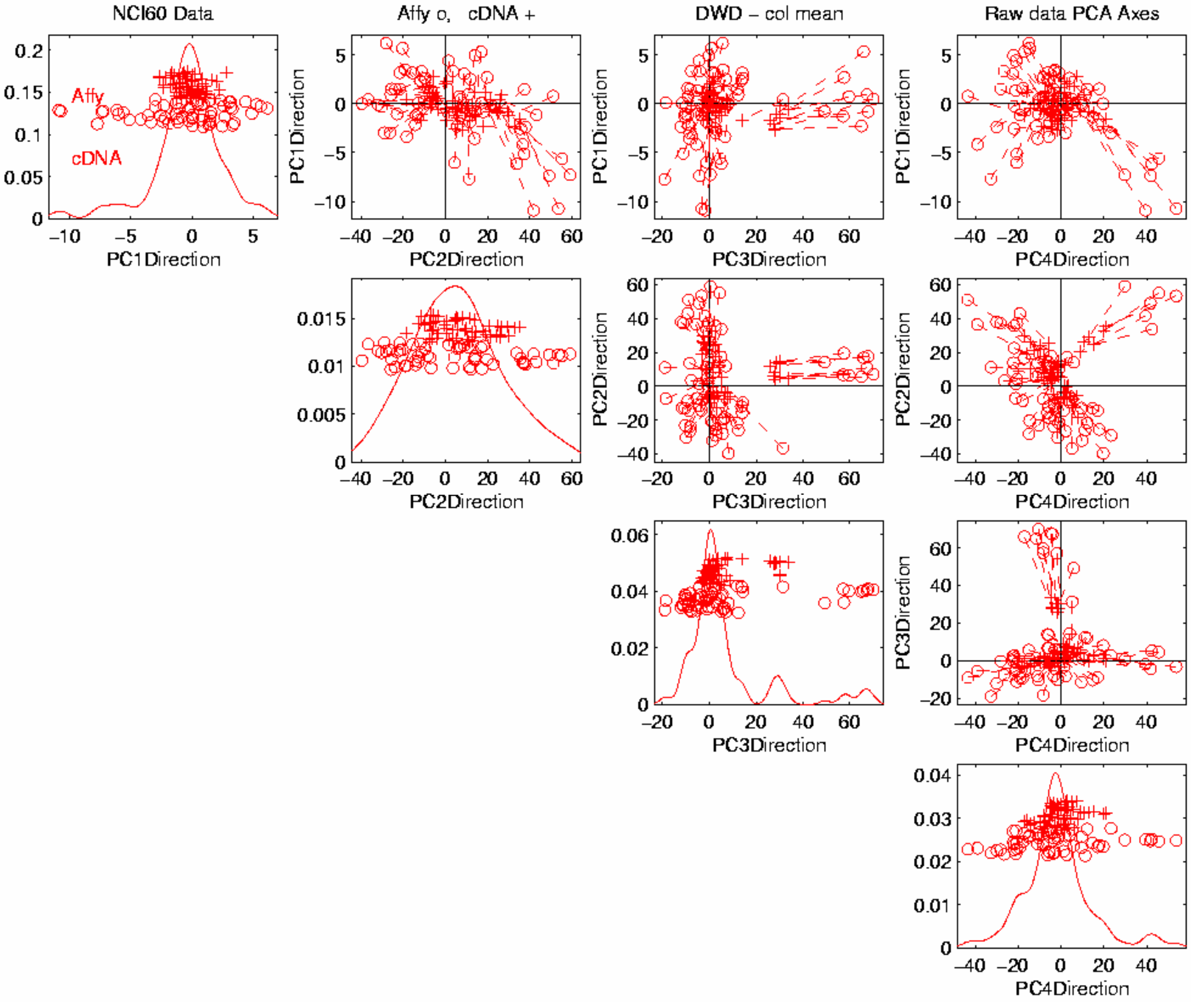
UNC Lineberger

NCI 60: Before & After Col. Mean Adj., Rescaled



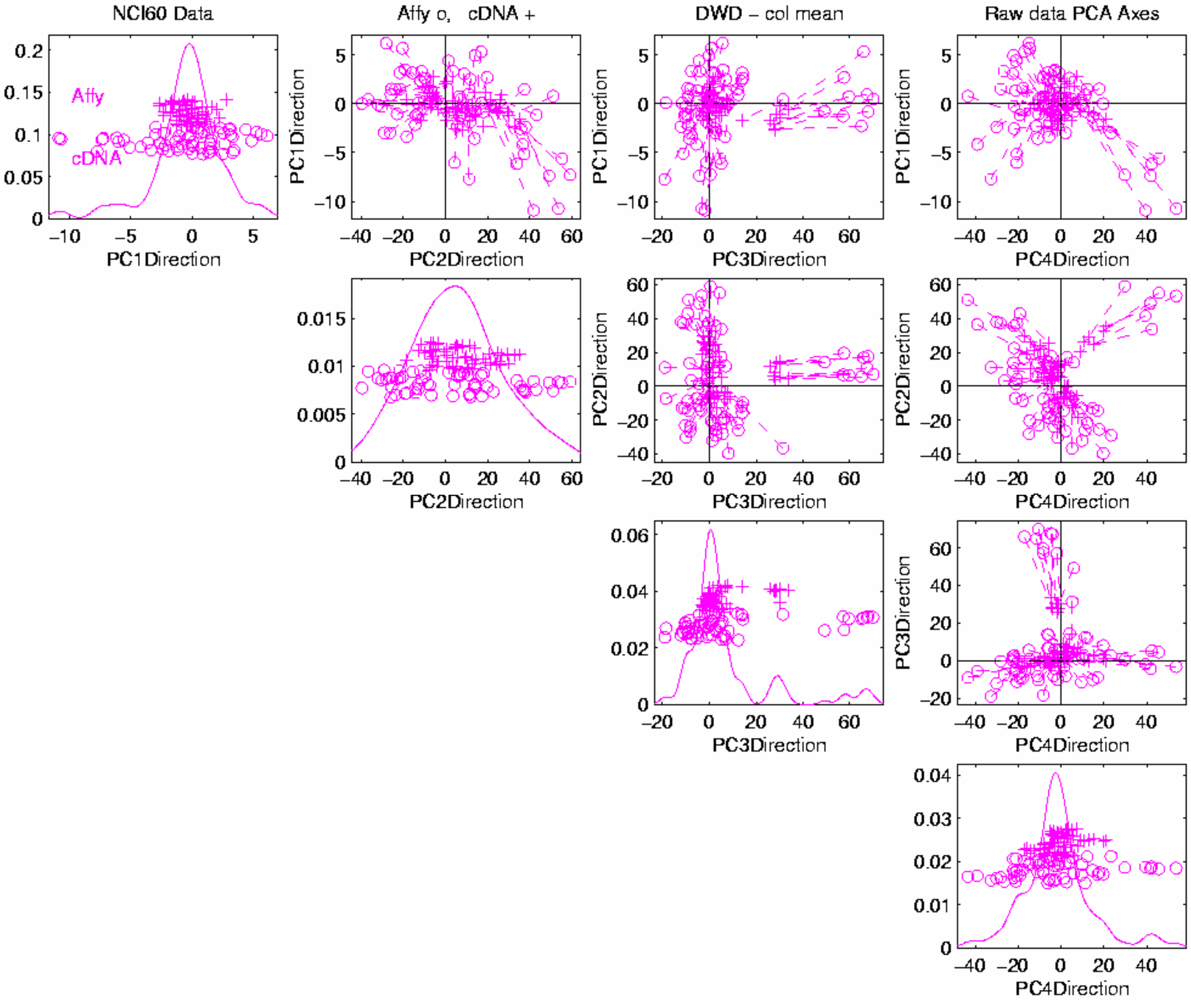
NCI 60: After DWD & Column Mean Adj.

UNC Lineberger



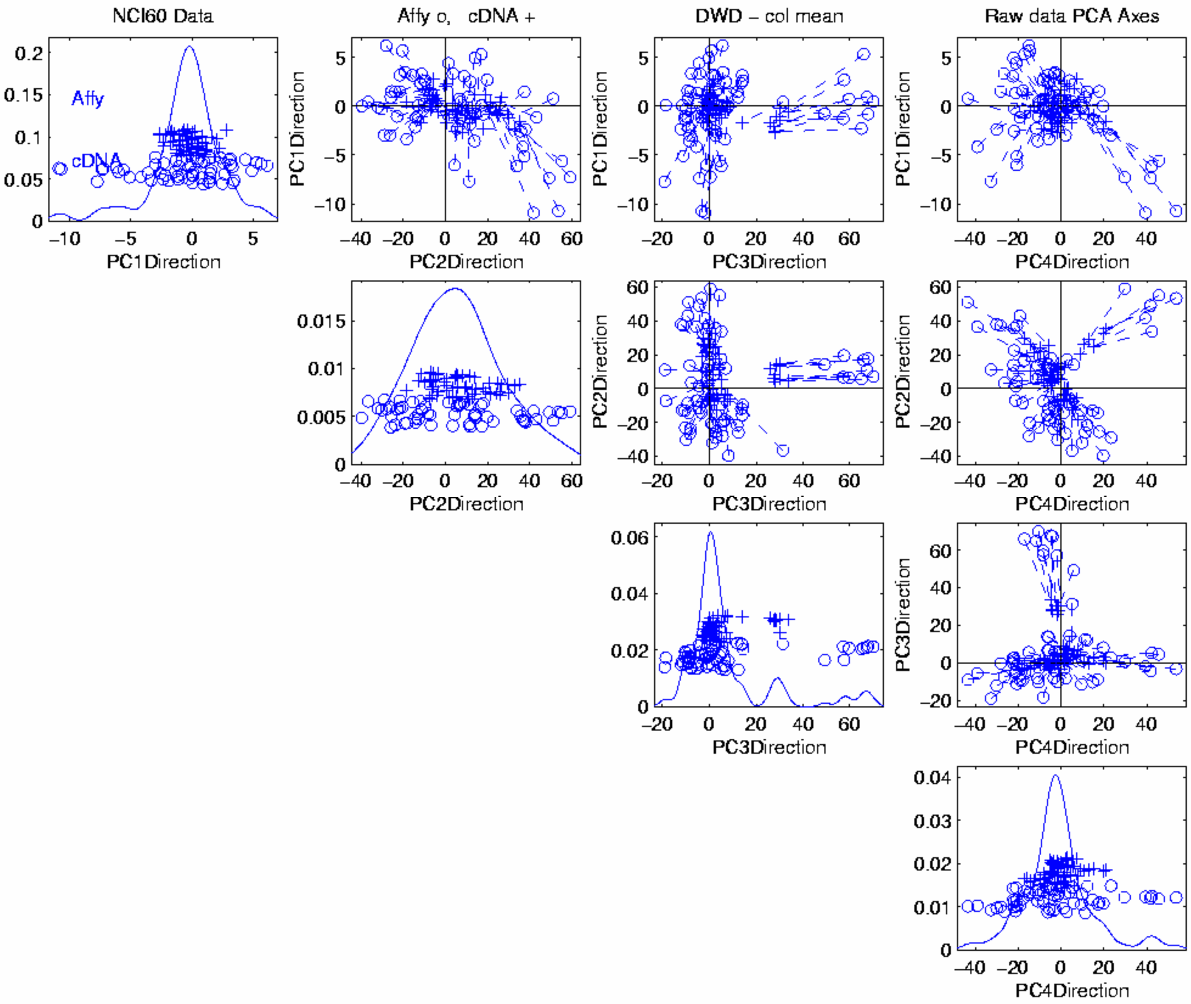
UNC Lineberger

NCI 60: DWD & Column Mean Adjusted



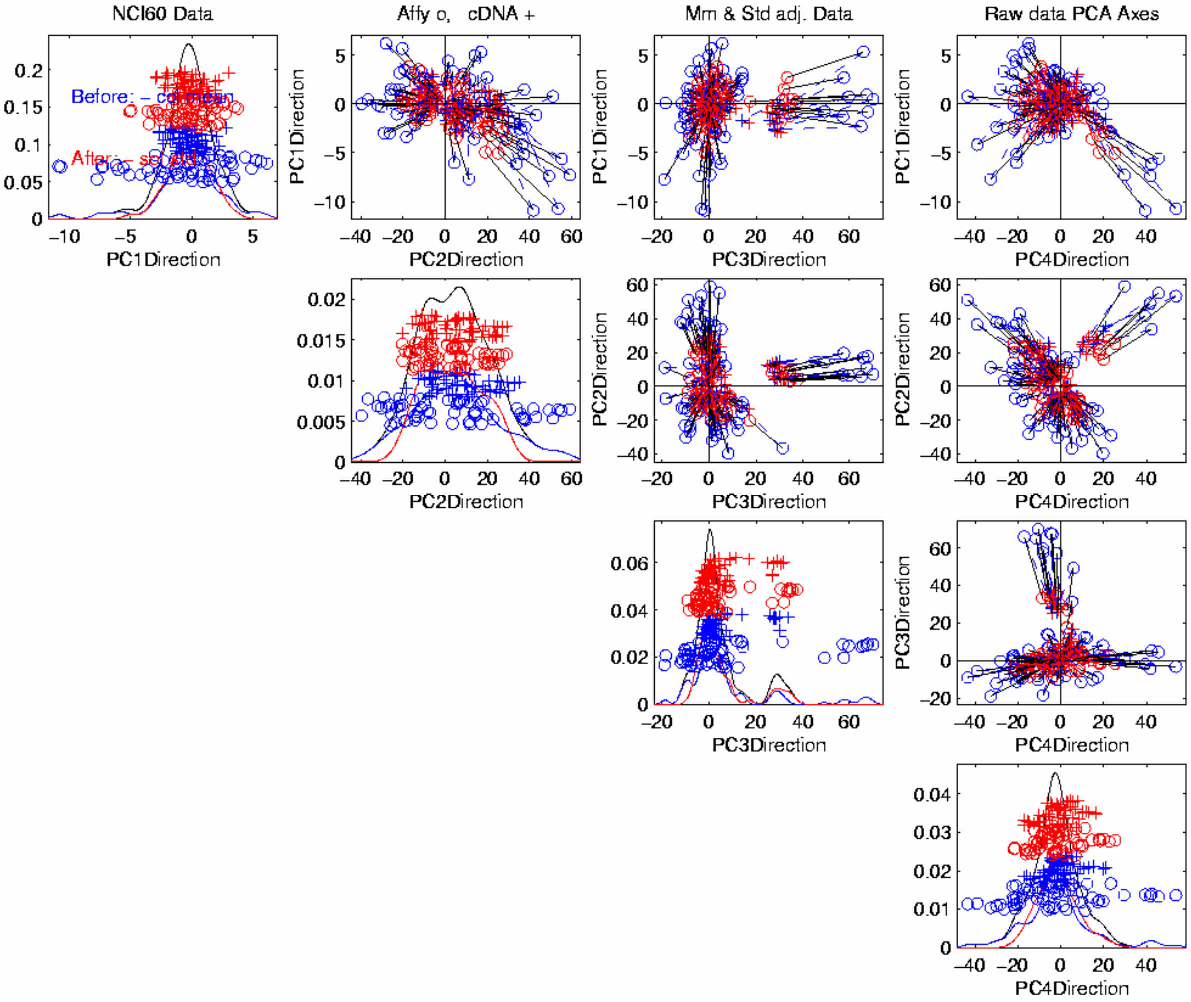
UNC Lineberger

NCI 60: Before Column Stand. Dev. Adjustment



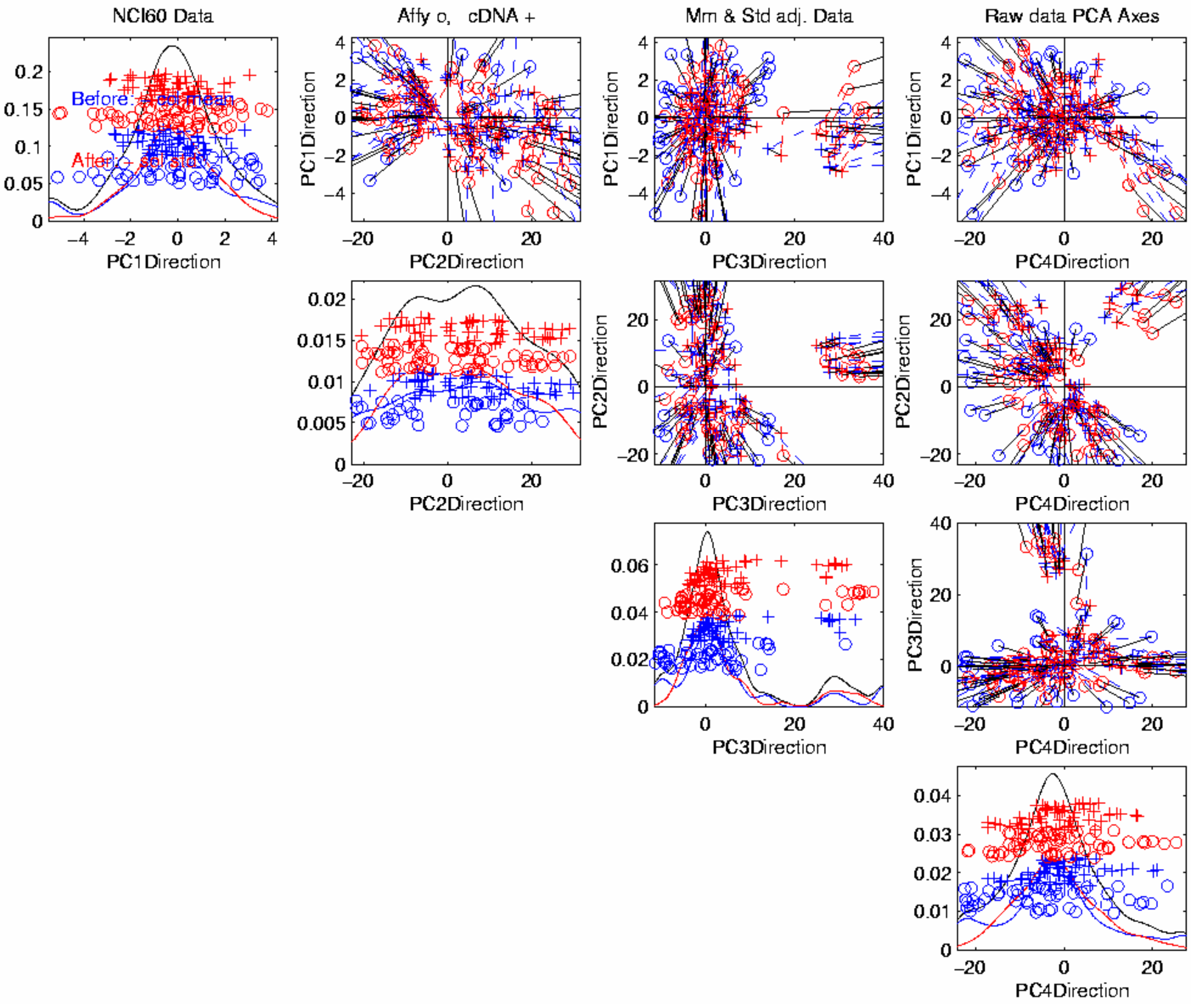
UNC Lineberger

NCI 60: Before and After Column S.D. Adjustment



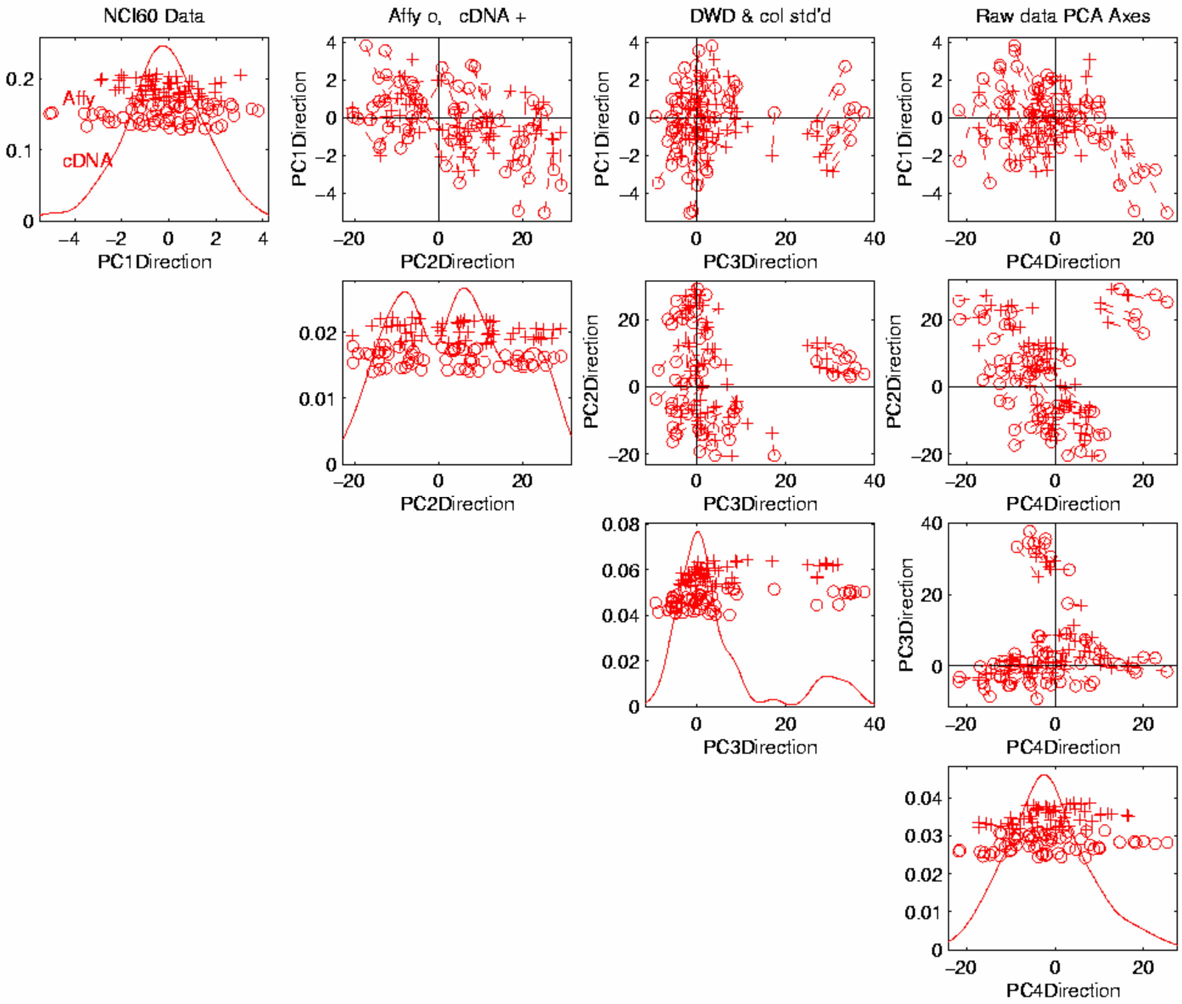
NCI 60: Before and After Col. S.D. Adj., Rescaled

UNC Lineberger



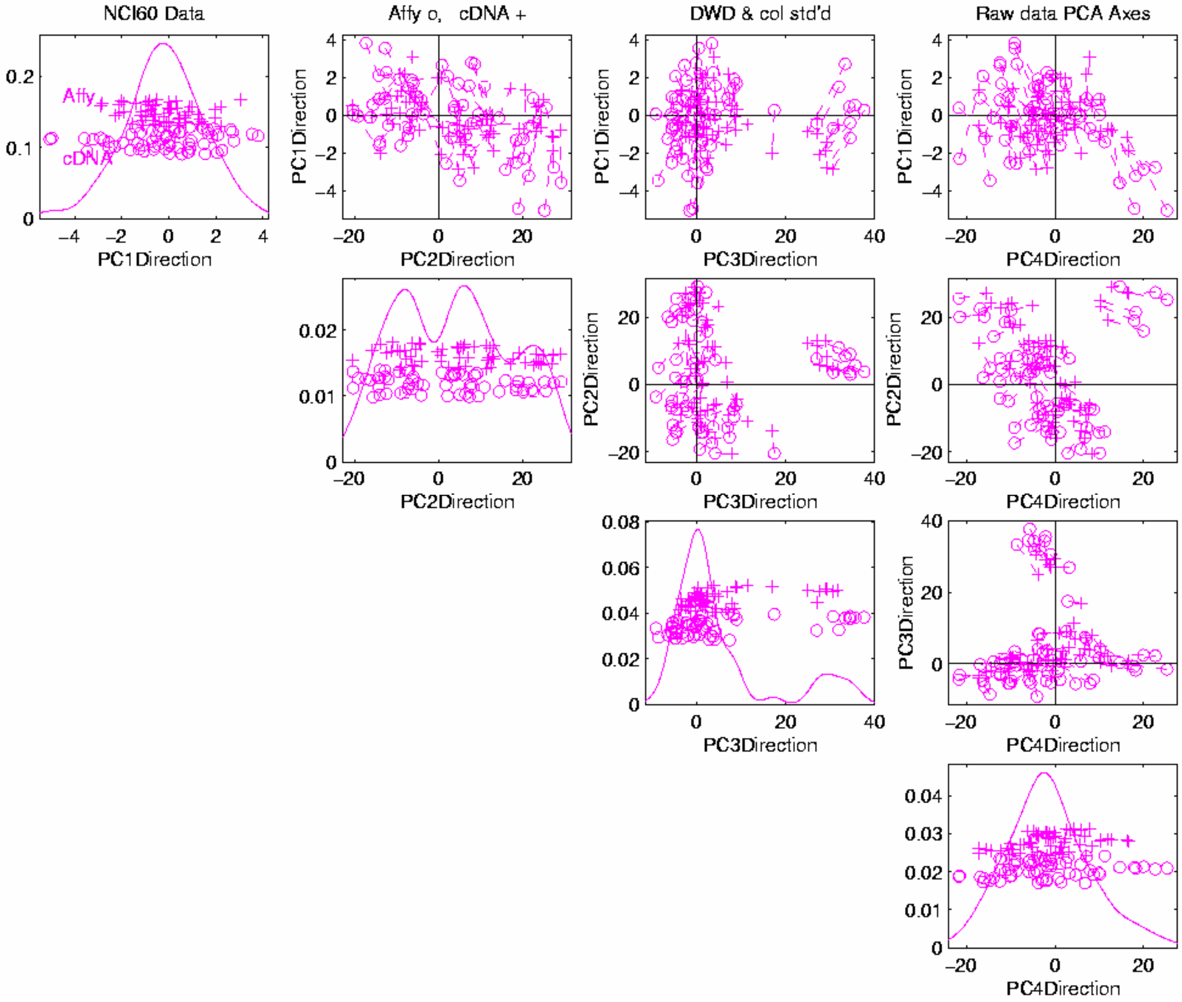
UNC Lineberger

NCI 60: After Column Stand. Dev. adjustment



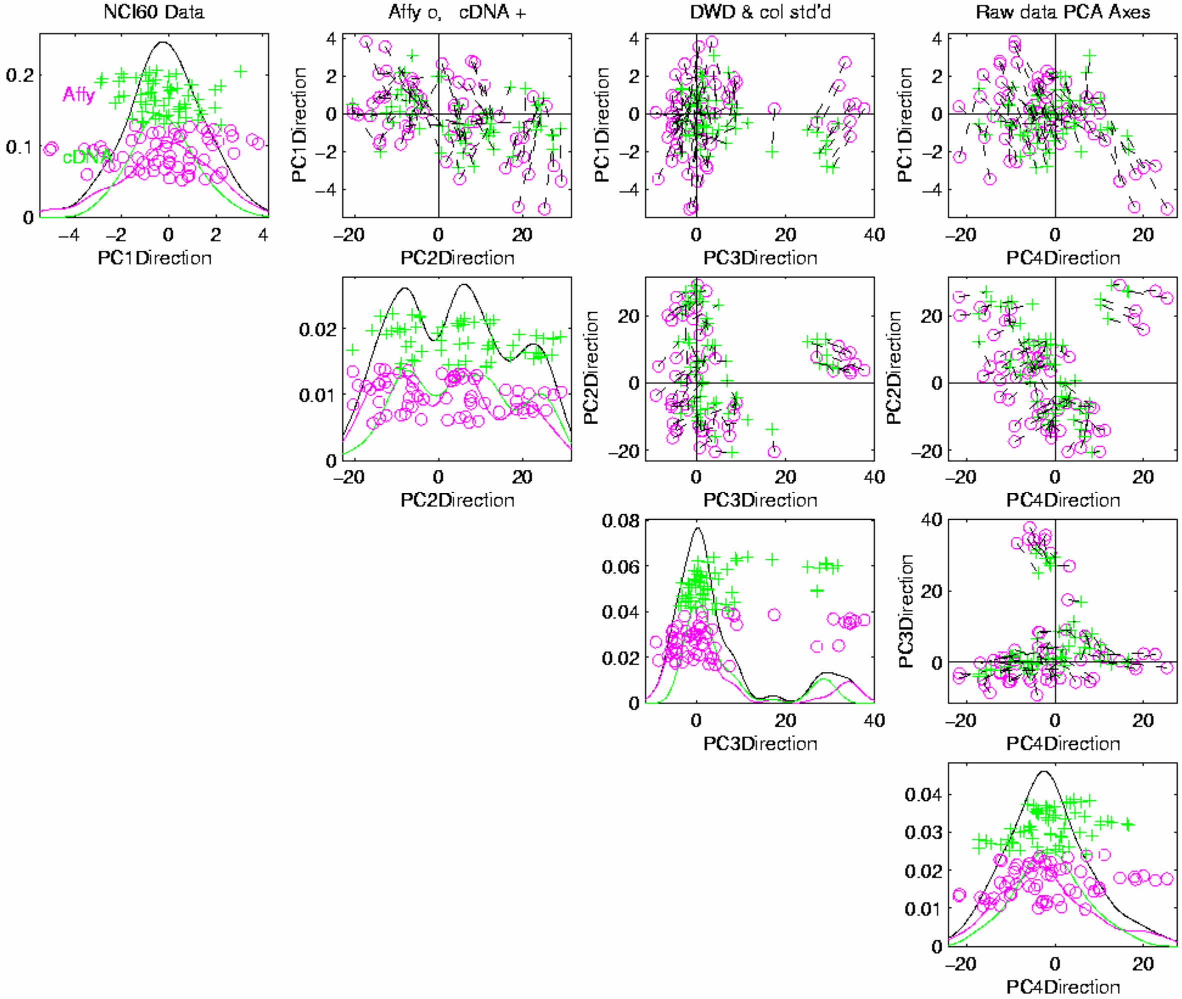
UNC Lineberger

NCI 60: Fully Adjusted Data



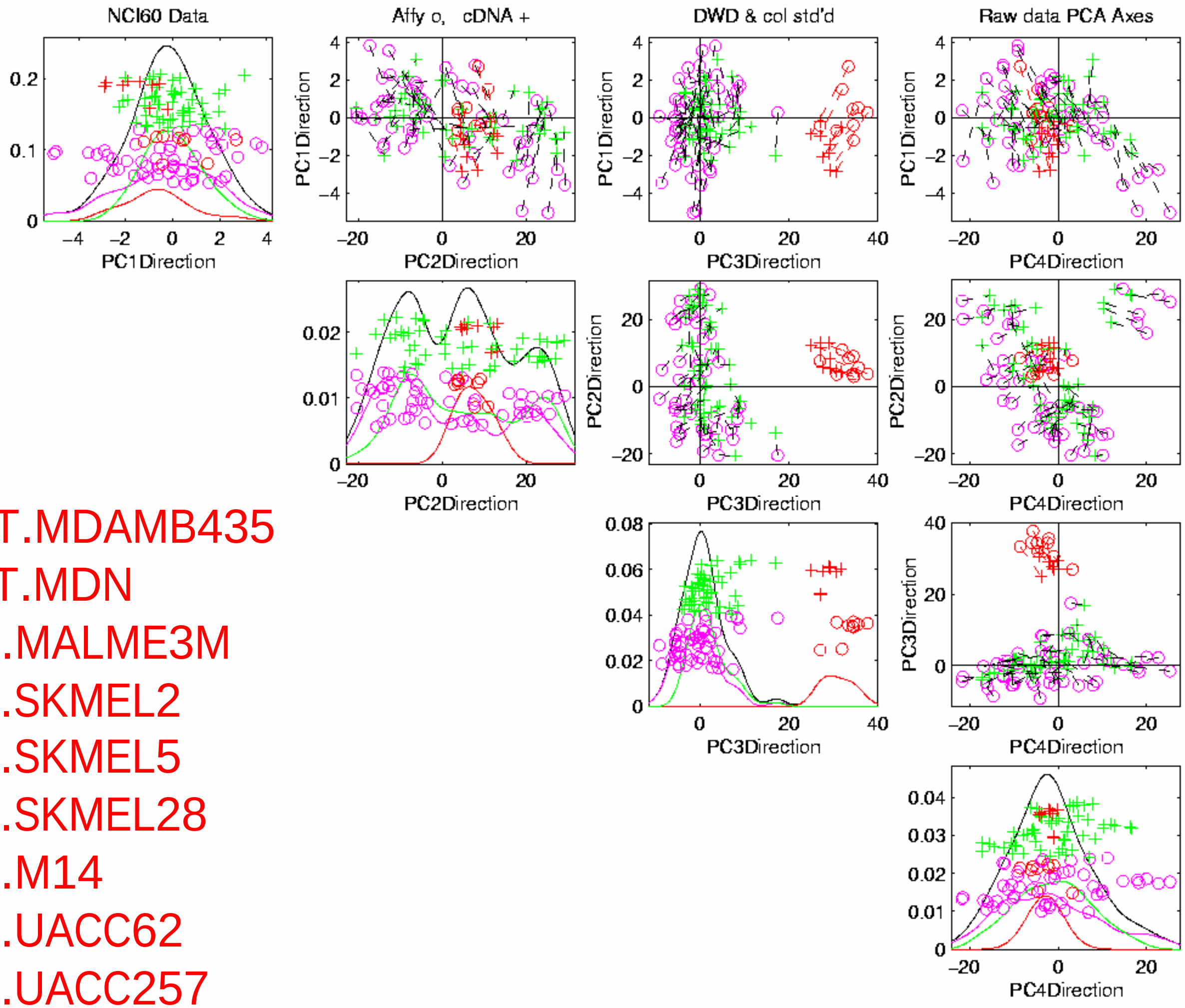
UNC Lineberger

NCI 60: Fully Adjusted Data, Platform Colored



UNC Lineberger

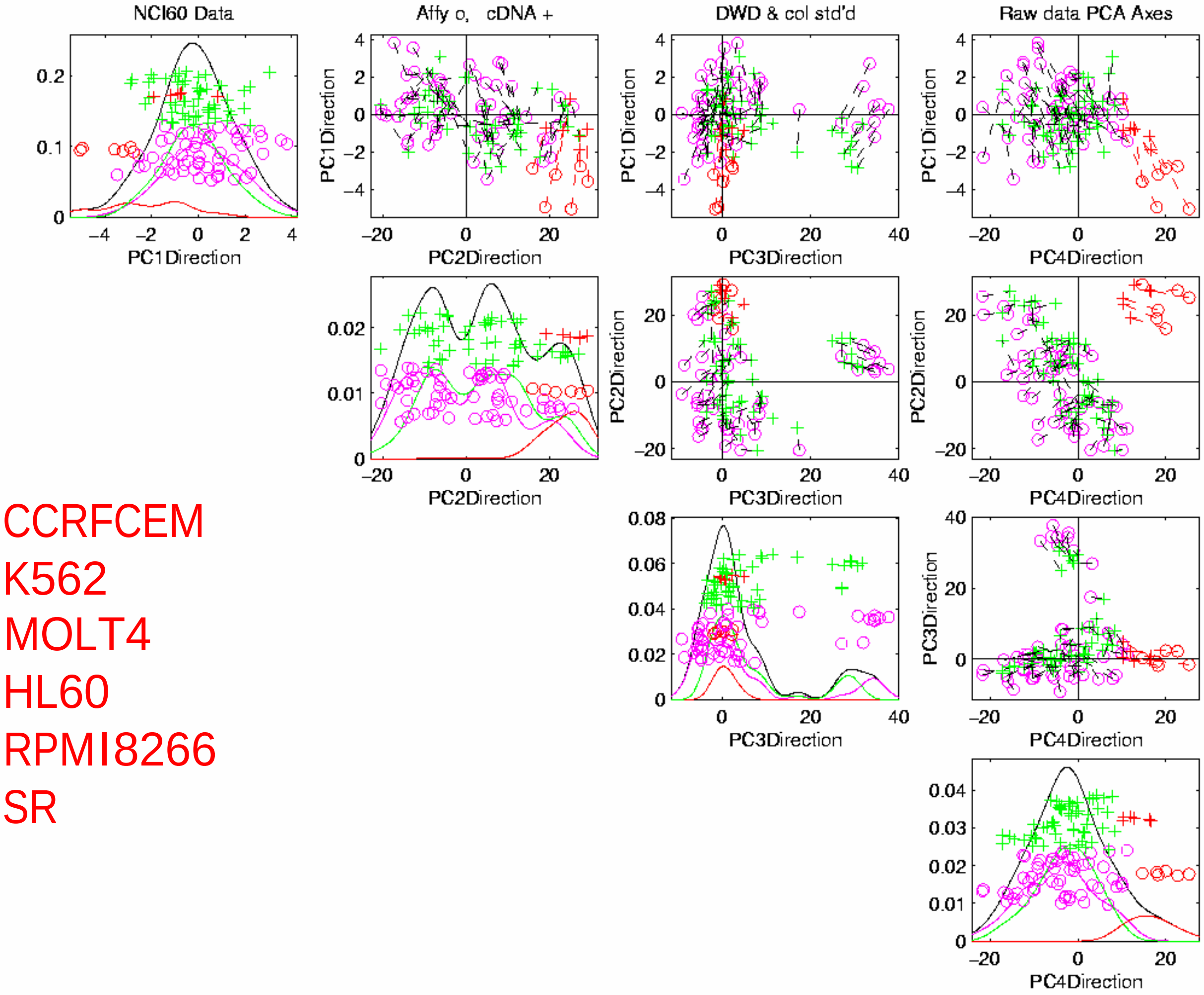
NCI 60: Fully Adjusted Data, Melanoma Cluster



BREAST.MDAMB435
BREAST.MDN
MELAN.MALME3M
MELAN.SKMEL2
MELAN.SKMEL5
MELAN.SKMEL28
MELAN.M14
MELAN.UACC62
MELAN.UACC257

UNC Lineberger

NCI 60: Fully Adjusted Data, Leukemia Cluster



LEUK.CCRFCEM
LEUK.K562
LEUK.MOLT4
LEUK.HL60
LEUK.RPMI8266
LEUK.SR

NCI 60 Adj: More Views

Internet Available:

http://genome.med.unc.edu:8080/caBIG/DWDindex.htm

Follow Link:

DWD Cross-Platform Adjustment of the NCI-60 Data

NCI 60 Controversy

- Can NCI 60 Data be normalized?
- Negative Indication:

  Kou, et al (2002) *Bioinformatics*, 18, 405-412.

  - Based on Gene by Gene Correlations

- Resolution:

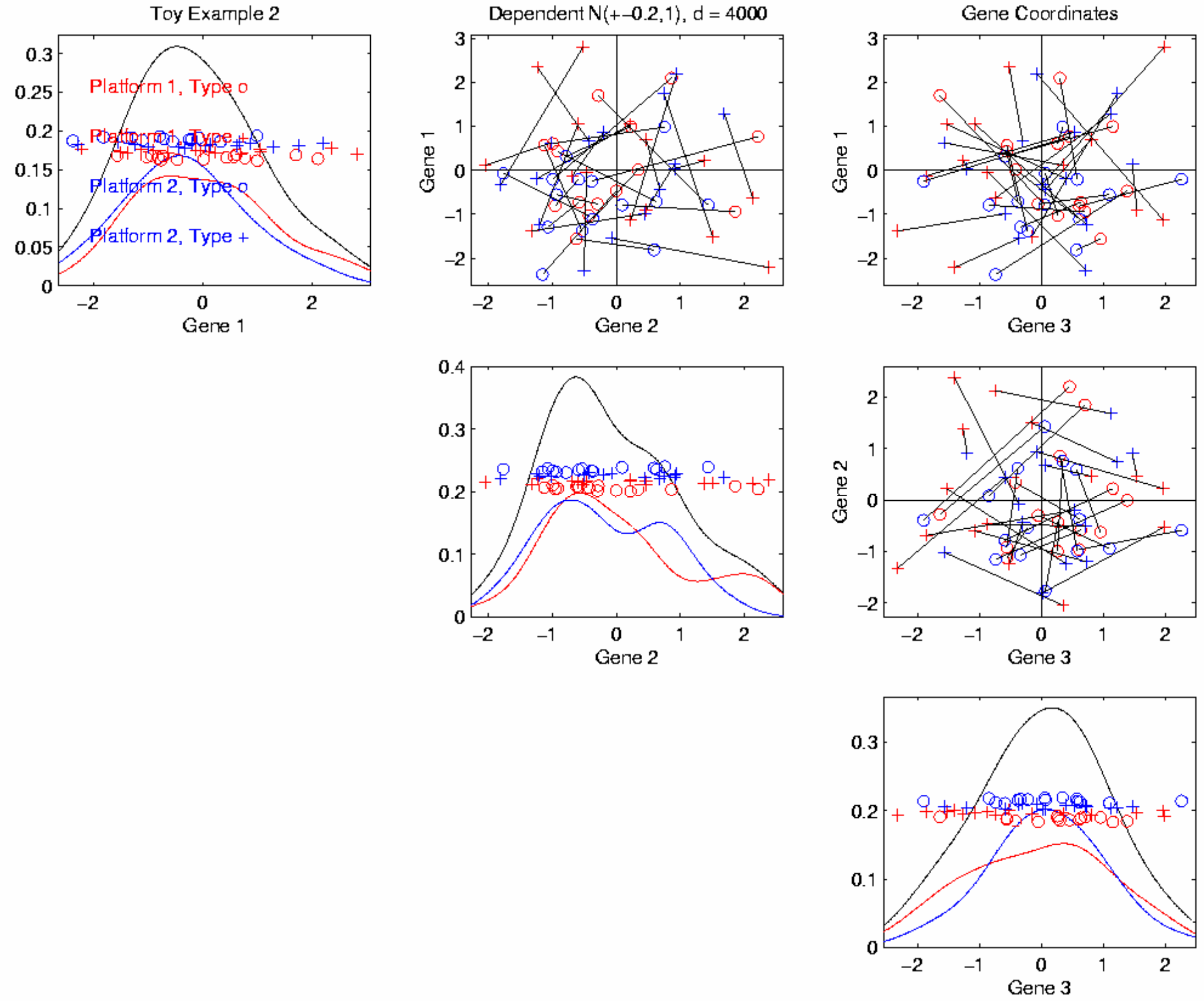
  Gene by Gene data view

  vs.

  Multivariate Data view

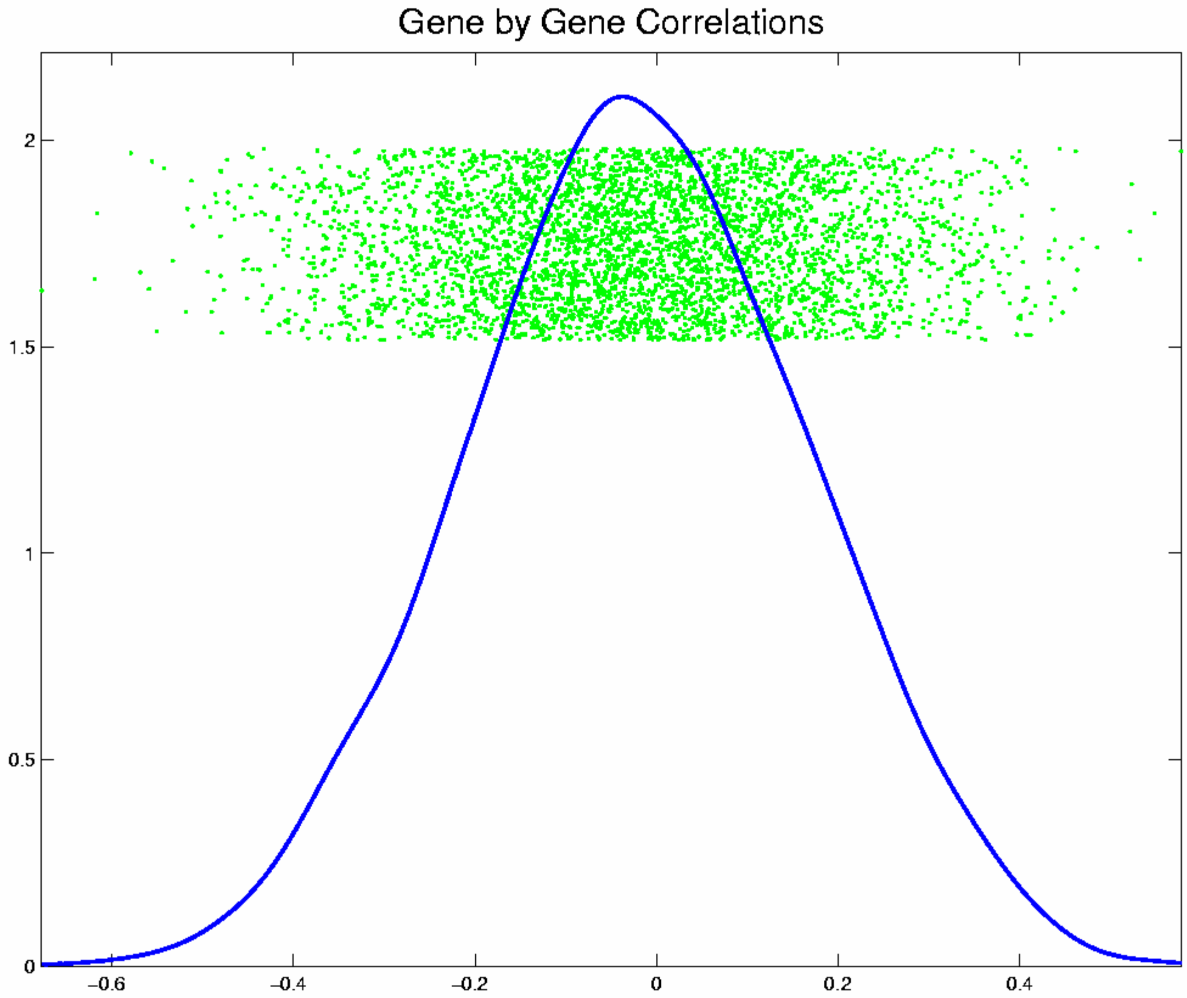
UNC Lineberger

Resolution of Paradox: Toy Data, Gene View



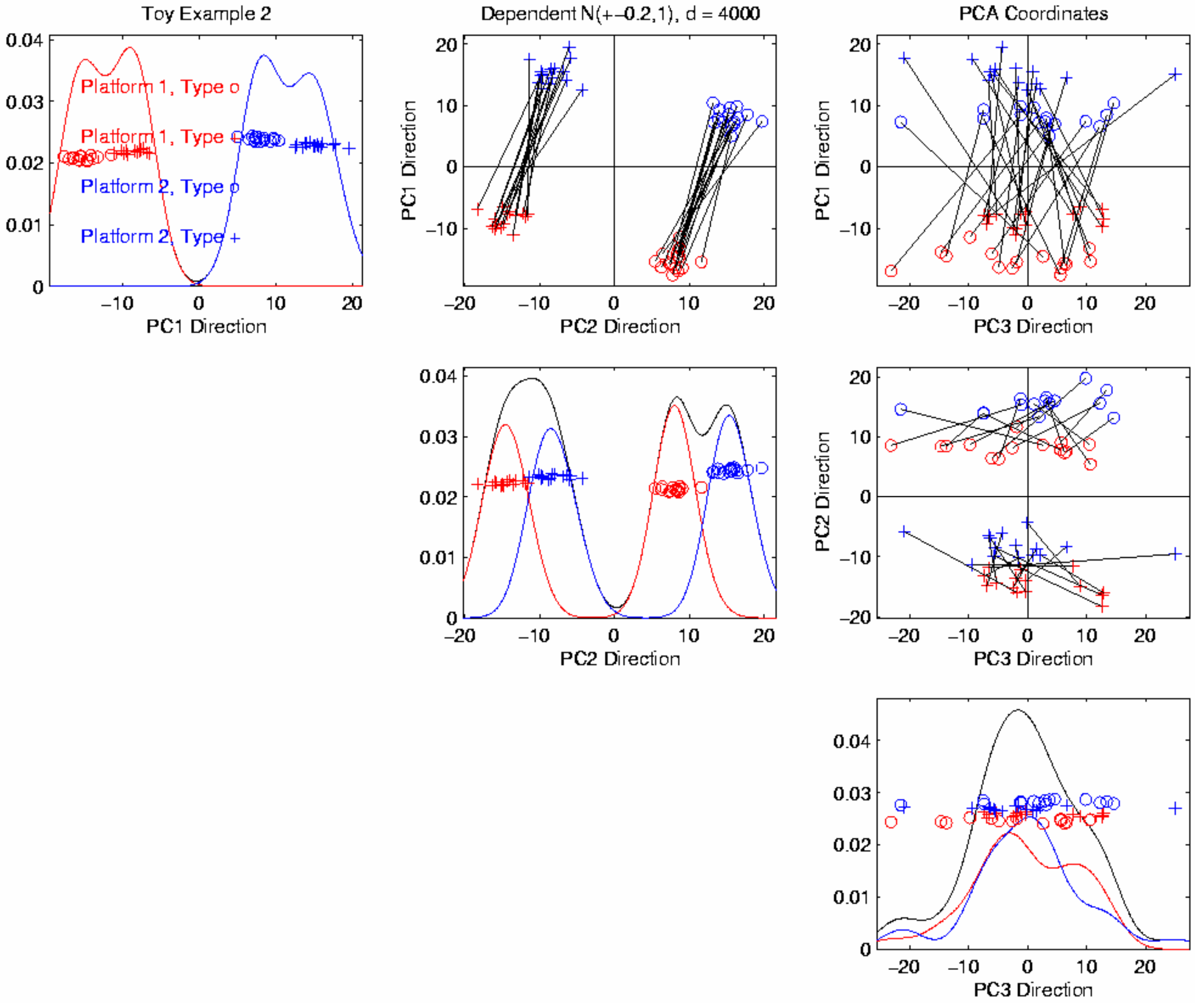
UNC Lineberger

Resolution: Correlations suggest “no chance”

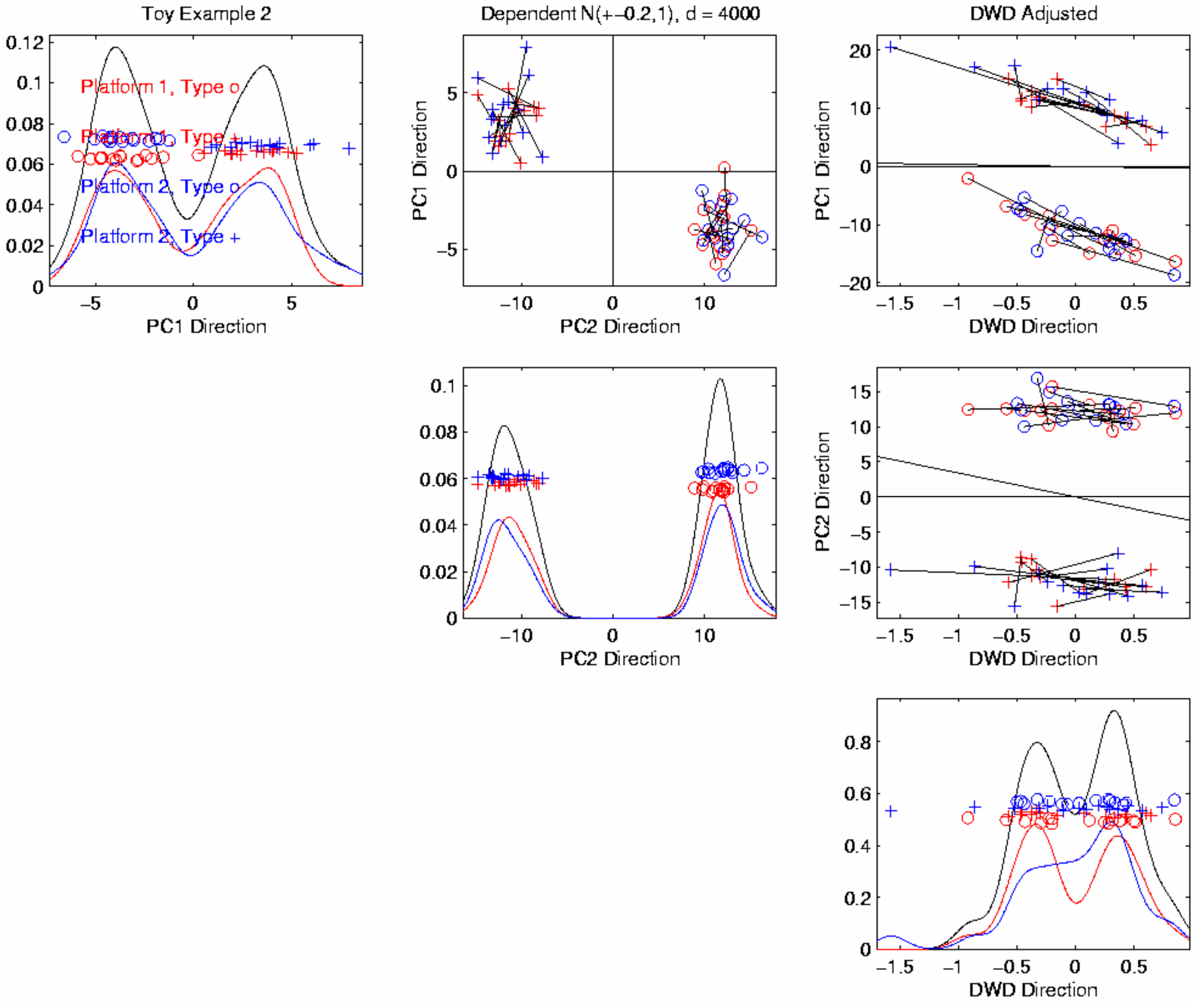


UNC Lineberger

Resolution: Toy Data, PCA View

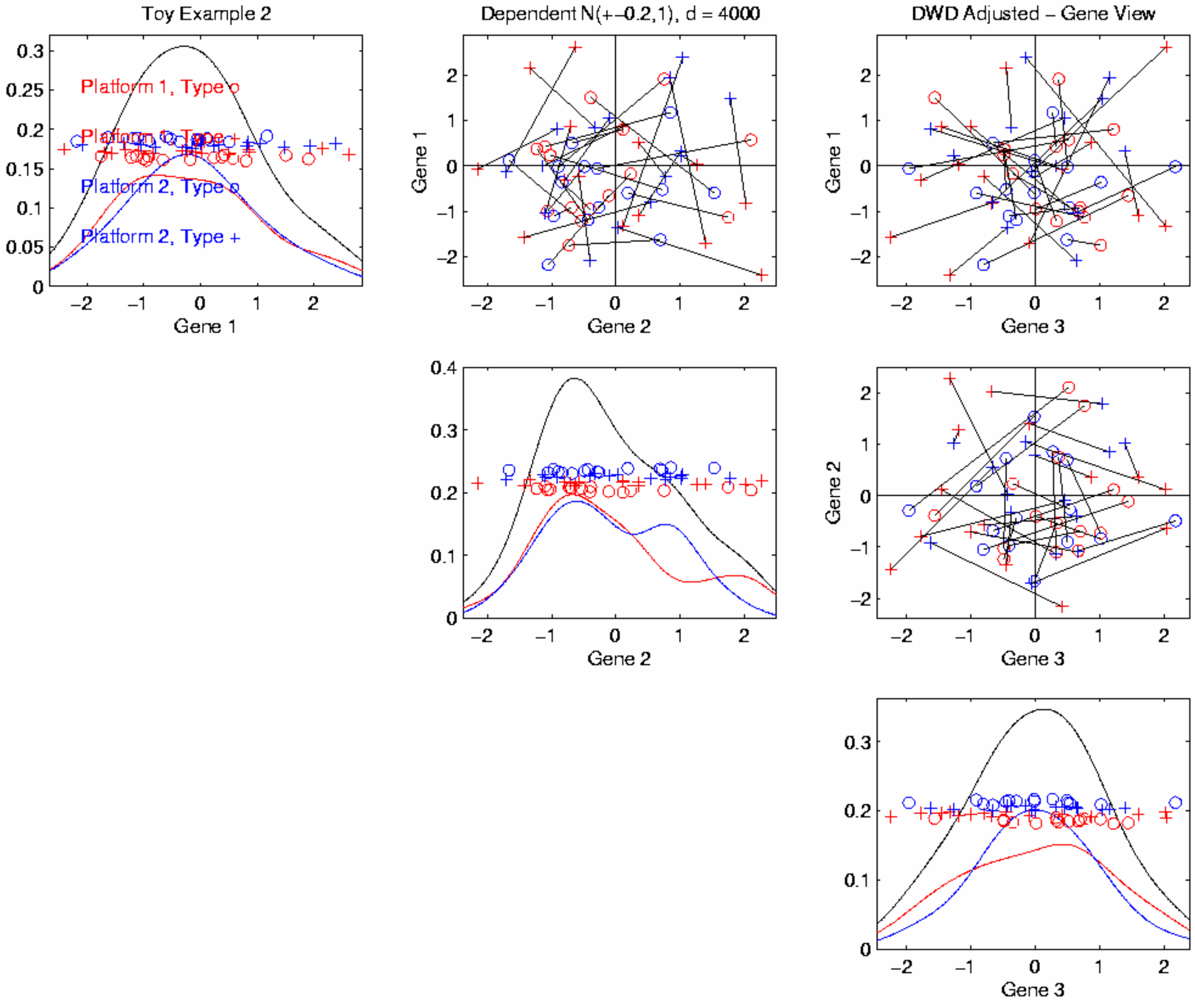


Resolution: DWD Adjusted



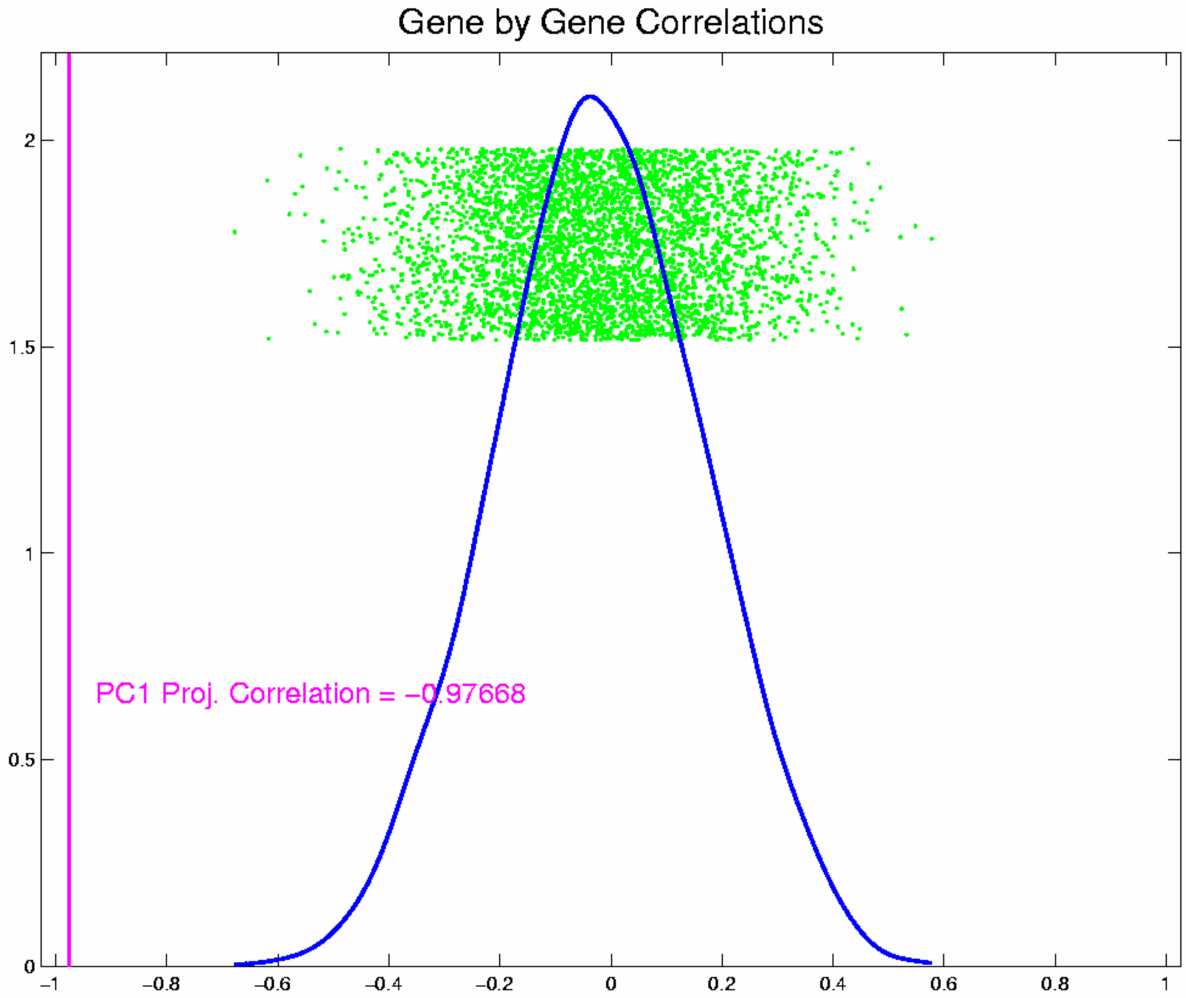
UNC Lineberger

Resolution: DWD Adjusted, Gene view



UNC Lineberger

Resolution: Correlations & PC1 Projection Correl'n



Needed final verification of Cross-platform Normal'n

- Is statistical power actually improved?

- A preliminary suggestion:
  - From C. Perou and J. Parker
  - DWD combined data across platforms
  - Split data into biological sub-classes
  - Got improved CV prediction of 5 year outcome
  - Suggests importance of "differing cancer types"

Careful about limitations

- Important Requirements:
  - All biological subtypes represented in all groups
  - Common gene sets
  - No missings
- Current state of the method:
  - No common samples
- Interested in prioritizing work on these?

Become an adopter!

Return to Key Philosophical Point

- Competing Paradigms:
  - Visually: what do we look at?
  - Conceptually: how do we think?

Gene by Gene

vs.

Multivariate “point cloud”

- Have illustrated power of multivariate concept

Website: Details & More Views

DWD caBIG Web Page:

http://genome.med.unc.edu:8080/caBIG/DWDindex.htm

- Many more “steps”
- Also Clustered Tree View Heat Map Views